(1)

ARI Research Note 90-79

Optimizing the Long-Term Retention of Skills: Structural and Analytic Approaches to Skill Maintenance

Alice F. Healy, K. Anders Ericsson, and Lyle E. Bourne, Jr.

University of Colorado

for

Contracting Officer's Representative
Michael Drillings

Office of Basic Research
Michael Kaplan, Director

July 1990

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

University of Colorado

Technical review by

George W. Lawton

DTIC
COPY
INSPECTED
6

Accesion For	
NTIS CRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY -- 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE --	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 153-0638	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 90-79
6a. NAME OF PERFORMING ORGANIZATION University of Colorado 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code) Campus Box B-19 Boulder, CO 80309	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600
8a. NAME OF FUNDING / SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-BR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-86-K-0155
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 61102B PROJECT NO.: 74F TASK NO.: N/A WORK UNIT ACCESSION NO.: N/A

11. TITLE (Include Security Classification)
Optimizing the Long-Term Retention of Skills: Structural and Analytic Approaches to Skill Maintenance

12. PERSONAL AUTHOR(S)
Healy, Alice F.; Ericsson, K. Anders; and Bourne, Lyle E., Jr.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 87/08 TO 88/08	1990, July	152

16. SUPPLEMENTARY NOTATION
Contracting Officer's Representative, Michael Drillings

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Skills, Skill maintenance, Memory structure, Training

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
This research program seeks to identify the characteristics of knowledge and skill most resistant to decay due to disuse. The program is divided into analytic and structural approaches. We performed two types of research to investigate skill retention and maintenance using the analytic approach. The first investigated different laboratory analogues of component military skills; the second investigated parallel natural skills learned by the college population during their prior education. We have developed five laboratory methodologies and completed experimental studies using each of them. We have also identified four natural skills and gathered long-term retention data for each of these skills. For the structural approach, we designed an experimental paradigm that allows us to assess the detailed encoding of new knowledge at presentation and at delay using verbal report techniques and chronometric measurement of retrieval components. Several studies of retention of vocabulary items have been completed within this paradigm.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Michael Drillings	22b. TELEPHONE (Include Area Code) (202) 274-8641 22c. OFFICE SYMBOL PERI-BR

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

OPTIMIZING THE LONG-TERM RETENTION OF SKILLS: STRUCTURAL AND ANALYTIC APPROACHES TO SKILL MAINTENANCE

EXECUTIVE SUMMARY

This research program seeks to identify the characteristics of knowledge and skill that are most resistant to decay due to disuse. Our research can be divided into two complementary parts. The first part is concerned with experimental analysis of factors influencing and improving retention of skill components. The second part is concerned with analysis and assessment of the structure of acquired memory and skills and how to monitor differential retention of components. The eventual goal of both parts is to be able to make relevant recommendations about training routines for long-term skill maintenance.

A new line of investigation, involving both the analytic and structural approaches, has recently begun consequent to the arrival of three Army tank simulators. This effort is concerned with the study of complex military skills. Extensive training of two subjects has been completed with the simulators.

The analytic approach. We have developed two lines of research for investigating skill retention and maintenance using the analytic approach. The first line of research involves investigating different laboratory analogues of component skills of electronic technicians. The second complementary line of research involves investigating parallel natural skills learned by the college population during their prior education.

We have developed five laboratory methodologies, and we have completed several investigations for each of them. The laboratory tasks involve (a) target detection, (b) data entry, (c) learning logical rules involved in circuit design, (d) memory for numerical calculations, and (e) temporal, spatial, and item components of memory for lists. We have also identified the following four natural skills and have completed investigations for each of them: (a) mental multiplication, (b) algebra, (c) data entry, and (d) temporal, spatial, and item components of memory for class schedules.

In studies of the data entry task we found that response latencies were significantly faster for old blocks of numbers responded to one month earlier than new blocks not seen previously. Further, for one subject given intensive training, we found improvements rather than losses in speed and accuracy after a 14-month retention interval. In studies of mental multiplication, we found improvements in speed and accuracy as a function of practice and poorer performance for more difficult problems (those involving larger numbers). In an investigation involving the long-term training of two subjects, we found that multiplication operations became more automatic with training; that is, as practice increased there was a smaller difference in speed and accuracy between easy and difficult problems. The first subject was retested after

retention intervals of 3 and 7 months and showed essentially no forgetting of this skill. The aim of our current work is to assess the extent to which automaticity is related to long-term retention of the multiplication skill.

The structural approach. We have designed an experimental paradigm that allows us to assess the detailed encoding of new knowledge at presentation and at delay using verbal report techniques and chronometric measurement of retrieval components. Two large-scale studies of retention of vocabulary items have been completed, in which subjects have been instructed to use the keyword method with supplied keywords. Subjects were assessed at three different retrieval tasks: the full retrieval task, the keyword task, and the image retrieval task. We found that the full retrieval task was slower than either of the component tasks and the keyword component showed better retention than the image component. We are currently conducting experiments to determine whether vocabulary items learned by the keyword method continue to be mediated by the keyword and image components with increased practice.

ANNUAL INTERIM REPORT FOR THE PERIOD AUGUST 12, 1987, TO AUGUST 11, 1988

In December three of us (Alice F. Healy, Lyle E. Bourne, Jr., and Robert J. Crutcher) visited the A.R.I. Field Unit at Fort Knox, Kentucky. We presented a briefing to the researchers there about the goals and accomplishments of our project. We also learned from them about the needs and interests of the Army. As a direct result of this visit, we plan to investigate in the future the more complex skills involved in tank gunnery and we will try to incorporate relevant task properties and components into the simpler skills we are already investigating to increase the relevance of our research to the skills learned by Army personnel. These efforts have been facilitated by our receipt in April of three TopGun tank simulators, on loan from the ARI research unit at Fort Knox, and by our recent completion of the extensive training of two subjects with these simulators. The details of the studies we propose to conduct with these tank simulators will be provided in a contract renewal proposal which we plan to submit in September.

In January eight of us (Alice F. Healy, Lyle E. Bourne, Jr., Robert J. Crutcher, David W. Fendrich, William Wittman, Lori Meiskey, Antoinette Gesi, and Robert Frick) met with three members of the A.R.I. staff (Michael Kaplan, Judith Orasanu, and Steven Goldberg) in Boulder to review our progress and discuss plans for future research, including those involving more complex tasks, as mentioned above. Further, two of us (Lyle E. Bourne, Jr. and K. Anders Ericsson) participated in the in-progress review meeting which met on March 15-17 in Champaign-Urbana, Illinois. At that meeting we summarized the progress on our project, with a focus on our work in the analytic approach involving the data entry and multiplication skills (see Appendix A) and our work in the structural approach involving the keyword method.

The Analytic Approach

We have made further progress in our testing of both the laboratory skills and the natural skills which we began in the first year.

Laboratory Skills

Target detection. Our initial work on target detection is summarized in a manuscript recently submitted for publication by Alice F. Healy, David W. Fendrich, and Janet D. Proctor. This manuscript, entitled "The effects of training on letter detection," is appended to the present report (see Appendix B).

In addition, we have completed training 24 subjects in a new experimental investigation of target detection. Each subject was trained for four one-hour sessions, half with a consistent-mapping procedure and half with a varied-mapping procedure. Surprisingly, we found little difference between the two types of training procedures, and both groups of subjects showed evidence of increases in the degree of automaticity of their traget detection skill. Further, both groups of subjects responded significantly more accurately, rapidly, and automatically to targets used during training than to new targets shown only at the end of training. In a posttest involving letter detection in prose, we found a large word frequency disadvantage (i.e., subjects showed poor performance on detecting letters in the very common word the). This result,

coupled with our earlier finding that the word frequency disadvantage is eliminated after exposure to a prose letter-detection task, indicates that extensive letter-level processing is not sufficient to eliminate the word frequency disadvantage. Rather, it seems that this effect is sensitive to word-level processes alone. However, contrary to this hypothesis, we found that subjects given consistent mapping training with the target H were significantly more accurate in detecting H during the passage letter-detection task (presumably because of enhanced letter-level processing) and, most crucially, the word frequency disadvantage was smaller for the subjects trained with H than for the subjects trained with other letters.

After a delay interval of approximately eight months, we recalled and retested these subjects. All but two subjects provided data for this retention test. In this test we found significant but not complete forgetting of the detection skill, and the degree of forgetting was not affected by the type of training (consistent or varied) administered. Nevertheless, subjects continued to respond more quickly, accurately, and automatically to the targets on which they had received extensive training relative to the targets shown only at the end of training. In addition, although these subjects were less accurate and slower at the retention test than at the final acquisition test on detection of letters in displays of random characters, their performance improved over the eight-month retention interval on detection of letters in prose. This improvement is consistent with our earlier finding that the word frequency disadvantage was reduced or eliminated after exposure to a pretest prose letter detection task. Our present finding, though consistent with our previous work, is particularly noteworthy because of the considerable delay (eight months) between the pretest and retention test.

We retested our single subject given extensive long-term training after an additional six-month retention interval. The subject was tested with the trained target as well as with a new target. In both cases retention of the skill was essentially perfect, with only a small difference between the old and new targets. In addition, we completed long-term training of three new subjects. In two of these cases, the subjects were administered varied-mapping training, instead of the consistent-mapping procedure used for the original subject given long-term training. The third subject was given training identical to that administered to our first subject. The varied-mapping subjects completed 12 training sessions followed by 2 final test sessions in which we examined transfer both to new targets and to new distractor characters. Although these subjects during the initial training sessions showed improvements in the speed and accuracy of target detection comparable in magnitude to those exhibited by the subjects trained less extensively with the same procedure, further training beyond the initial sessions led to little further improvements in the detection skill. In contrast, the third subject, who was given consistent-mapping practice like the original subject, showed large improvements throughout the entire course of training. We plan on retesting these three subjects after a retention interval of approximately six months, with both original training and transfer tests. At present only one of these subjects has been retested.

We have also designed a multi-subject training and transfer target-detection experiment in which we will systematically vary the visual similarity of the targets to the distractors and to the filler characters. We plan to begin this experiment during the next year.

In collaboration with Janet Proctor of Auburn University, we have completed two experiments. These experiments follow up our finding that subjects given experience with detecting letters in a prose passage no longer exhibit the word frequency disadvantage typically found in the letter detection task. In these experiments we gave subjects strong and subtle hints that the target letter can be found in the word the and we collected retrospective verbal reports from the subjects about the strategies they employed. Our results led us to conclude that a strategy shift to looking for the word the is a likely factor in the reduction of the word frequency disadvantage but that other factors, perhaps perceptual, must also be involved. The results of the two initial experiments were reported at the annual meeting of the Southeastern Psychological Association held in March in New Orleans. A copy of that paper is appended to this report (see Appendix C). The paper, coauthored by Janet D. Proctor, Alice F. Healy, and David W. Fendrich, was entitled, "The disappearance of a word inferiority effect: Strategy shift or perceptual effect?" Most recently we have completed the testing of subjects in a follow-up experiment investigating alternative explanations for these effects. We are presently analyzing the data from this experiment.

Data entry. We have completed five new experiments with the data-entry procedure. In the first experiment, 24 subjects were tested in two sessions with a one-day interval separating them. In each session subjects were shown 30 blocks of 10 three-digit numbers, and they were required to type them on the keypad as quickly and accurately as possible. Half of the blocks shown during the second session were old (i.e., they had been shown during the first session) and half were new. We found a significant improvement in typing speed on the second day of training, and a significant advantage for the old blocks relative to the new blocks. This result replicates the major surprising finding from our original data entry study.

In the second experiment, 24 additional subjects were tested in the same manner except that the three-digit numbers were shown individually rather than in blocks. Further, on the second day subjects made old/new recognition decisions for each three-digit number immediately after typing in the number. We found that subjects did show a low but significant level of recognition for the old three digit numbers. Further, we found that they responded significantly more rapidly to the old that to the new numbers, but only when they correctly recognized those numbers.

The third experiment further explored the differences between massed and spaced repetitions of stimulus items and assessed retention of the stimuli after a one-month interval, both in terms of improved data entry speed and by means of explicit recognition ratings by the subjects. This experiment allowed us to test the hypothesis that memory for the stimuli is dependent on the subjects' repeating their motor responses to the stimuli rather than their perceptual encoding of the items. We found that subjects' recognition of the stimuli was enhanced but not dependent on repetition of the motor responses. Further, we found that there was an improvement in typing speed for the stimuli seen one month previously, but only when those stimuli were explicitly recognized by the subjects. In addition, we were able to analyze the typing responses in terms of the individual keystrokes. This analysis revealed that the largest facilitation in typing speed for the old items occurred on the first of the three digits in a sequence.

The fourth experiment made use of two different configurations of the

keypad (one like that found on a typical computer console and one like that found on the telephone) in order to determine if subjects retain information about the motor sequence of key presses or information about the actual sequence of digits displayed. In addition, this experiment allowed us to assess the difficulty of transferring the data entry skill from one keypad configuration to the other, after a one-week retention interval. We found that overall the two configurations yielded equivalent performance and switching to a new configuration did not significantly impair performance. Most crucially, we also found that when the keypad configuration was switched, the advantage for the old items occurred when either the same digits were repeated or the same motor pattern was repeated. Hence, the subjects' long-term memory representation of the digit sequences must contain both motor and perceptual information.

The fifth experiment also employed two different key configurations, in this case the keypad and the horizontal linear array of digits at the top of the standard keyboard. Like the third experiment this study examined long-term memory for the digit sequences both in terms of explicit recognition responses and in terms of the speed and accuracy of entering the digits. Also, like the fourth experiment, we assessed the extent to which a match in the method of responding to the stimuli at study and at test influenced the two measures of memory. An additional new condition allowed us to compare the effects on long-term retention of reading and entering the stimuli to the effects of just reading alone. We are presently in the process of analyzing the data from this study.

Memory for numerical calculations. We completed a manuscript describing our initial work on memory for numerical calculations. This manuscript is appended to the present report (see Appendix D). The title of the manuscript is "Cognitive operations and the generation effect." The coauthors are Robert J. Crutcher and Alice F. Healy.

In addition, we completed a third experiment in which we tested subjects' memory for simple multiplication problems. In this experiment we varied the method of answer production both at the time of study and at the time of test. Subjects either solved the problem in their heads or they used a calculator to solve the problem. We found a small effect of test appropriateness; that is, subjects showed better recognition of a problem if they used the same answer production method at test as they had used at study. However, this effect was only marginally significant and was overwhelmed by the large effect of the method of answer production used at study. Subjects' recognition of a problem was much better if they had mentally computed the answer rather than using a calculator. Hence, this experiment provided further support for the importance of internal cognitive operations in enhancing memory performance.

In addition, an undergraduate honors student, Lauarlyn Andrews, completed a thesis under our direction this year. The thesis involved two experiments further exploring the generation effect, in this case with verbal materials rather than with numerical calculations. These experiments provided support for our hypothesis that the basis for the generation advantage involves internal cognitive operations by the subjects.

Learning logical rules involved in circuit design. We have initiated a follow-up investigation of this skill. In this new experiment subjects were either given only geometric symbols standing for each rule, given meaningful labels (like "and" and "or") for each rule, or given both types of information.

Our aim is to determine whether the rule learning behavior will show a change in pattern when the nature of the rules is made more explicit. We completed testing subjects in the experiment and are now in the process of analyzing the data.

Temporal, spatial, and item components of memory for lists. An experiment following up our initial work on this task has been completed. The major change in procedure involved a difference in the spatial arrangement of the stimulus presentation display, thereby resulting in a difference in the spatial location information to be learned. Specifically, subjects were shown 18 words in two three-by-three matrices, rather than the vertical spatial array used previously. The new format was expected to facilitate the learning of spatial information and was thus expected to enhance performance on the spatial component relative to the item and temporal components. The analyses confirmed that acquisition and retention (across a six-week delay) were best for item information, followed by spatial information, and were worst for temporal information. In contrast, the previous experiment, which used the vertical spatial array, had demonstrated that performance was considerably worse for spatial than for temporal information. It was predicted that changing to a matrix array would provide more appropriate spatial cues with which to learn the spatial arrangement of the stimuli than employing the cues existing in a vertical array such as that used in the previous experiment. The results strongly support this hypothesis.

In the present experiment, retention intervals of one week and six weeks were compared. There were significant differences between the two intervals for all three types of information. Hence, these components of memory for lists did demonstrate substantial forgetting, in contrast to the minimal amount of forgetting we found for the target detection, data entry, and multiplication skills.

Natural Skills

Mental multiplication. We have completed long-term training in mental multiplication of two subjects. These subject were given 11 acquisition sessions with the keypad method of responding and a final 12th session with the oral method of responding. We found dramatic improvements in the multiplication skill, primarily in terms of speed of responding, because accuracy was essentially perfect. Differences in the speed of responding as a function of problem difficulty (i.e., the magnitude of multipliers) decreased to some extent as training progressed but were clearly evident even at the end of training, thus suggesting that the subjects had not yet reached the point of automatism in this skill.

We retested the first subject given long-term training in the multiplication skill after retention intervals of three months and seven months. She showed essentially no loss of either speed or accuracy on this task, despite the fact that she had shown substantial gains in performance during the 12 acquisition sessions. Hence, this skill resembles the target detection and data entry skills in its retention characteristics.

We have also designed a multi-subject experiment to assess the hypothesis that long-term retention of the multiplication skill is affected by the extent to which subjects have become automatic at the skill. This experiment will also determine whether training on this task is influenced by the order of the multipliers given to the subjects. Specifically, subjects will see each problem

in only one of the two possible orders (e.g., 2 X 3 or 3 X 2) during the acquisition phase of training but will see both orders during the retention phase. This experiment will enable us to determine whether only one or both orders need to be trained for successful long-term performance. A questionnaire assessing subjects' natural experience with multiplication will be administered to determine whether and how the skill is affected by experience with this task outside the laboratory.

Algebra skills. We completed the retesting of 15 students who were part of the initial group of students who took the introductory algebra class which we studied last year. These students had been given a multiple-choice test at the end of their course. The retesting took place after a retention interval of approximately six months. The retesting included both a multiple-choice algebra test like that given earlier and a questionnaire designed to assess how much training in algebra the students had received before the college algebra course and how much they had used algebra following the course. A stepwise regression was performed in which retest scores were predicted from the final algebra course grade, the score on the original multiple-choice test, and three variables derived from the questionnaire: number of semesters of math taken before the algebra course, whether or not the student was currently enrolled in a mathematics course, and how much the student had used algebra since completing the algebra course. Final course grade accounted for the most variance and was the only independent variable that made a significant contribution to predicting retest scores. Subjects' total scores on the retest were not significantly different from those on the original test taken at the end of the algebra course. However, on the retest, subjects showed a decrement in performance on three particular types of questions. In order to solve these questions, a particular rule had to be remembered (as opposed to remembering more general procedures, such as those involved in isolating a variable). The three types of questions which showed a decrement involved using the quadratic equation to solve a problem, using the rule for how to combine terms containing exponents, and using the rule to complete the square in a polynomial equation.

We also tested a new group of students in an introductory algebra class at the beginning and end of the course. We found significant improvement across the two testing sessions. We then retested, after a retention interval of approximately four months, some of these students. We are presently analyzing these retention data. We are testing the hypothesis that the information learned during the course, as opposed to the information already available at the beginning of the course, is most fragile and susceptible to forgetting during the retention interval. We are also examining whether the same types of information identified as showing some loss after a retention interval by students in the previous study show some loss by the students in this new study.

We reported our progress in studying the retention of algebra skills at the annual meeting of the Rocky Mountain Psychological Association held in April in Snowbird, Utah. A copy of that paper is appended to this report (see Appendix E). The title of the paper is "Long-term retention of algebra." The coauthors are Lori Meiskey, Alice F. Healy, Robert W. Ellingwood, and Lyle E. Bourne, Jr.

Temporal, spatial, and item components of memory for course schedules. We initiated a new study of this topic. In this new investigation we used a cued-recall procedure, rather than the questionnaire procedure employed in our pilot work. Specifically, three groups of sixteen subjects were tested with cues providing the when (the time of day), where (the building location), who

(the instructor), or what (the course title) of a course. All subjects were undergraduates in their third year of study at the University of Colorado. One group was questioned about their courses taken two semesters ago, the second group about their courses three semesters ago, and the third group about their courses four semesters ago. This initial testing provided a cross-sectional assessment of memory for this information. Our initial analyses indicate superior retention of item and spatial information relative to temporal information and information about the course instructor. Further, differences in recall were not found over the four types of cues. Additionally, recall performance was found to be positively related to subjects' ratings of course pleasantness and did not vary across gender.

We retested most of these subjects to obtain data appropriate for a longitudinal assessment of their memory for their course schedules. Preliminary analysis shows similar results to those found after original testing the previous semester. Overall recall accuracy was poorer, however, likely due in part to the longer delay since original learning.

Data entry. We retested once more (after an additional eight-month interval) the subject given extensive practice with the data entry skill in a natural job environment. Relative to the first retention test, we found no change in the error rate (which was close to the floor) but a significant improvement in the speed of responding. Hence, there is clearly no evidence of forgetting this skill over the long retention interval. A more detailed breakdown of performance as a function of block revealed some initial forgetting after the retention interval followed by rapid improvement across the 30 blocks of the one-hour session.

We plan on retesting once again the single subject given extensive natural training in data entry. This new test will involve transfer to a new keypad orientation like that found on a touch-tone telephone, rather than that found on a computer terminal or calculator.

The Structural Approach

In our earlier work, we developed a methodology combining cued recall and verbal reports to assess the differential decay of various components of vocabulary retention using the keyword method. Results for a 1-week and 1-month retention study suggested that the keyword component of the task was much less likely to decay than the image component. We completed a new study to replicate and extend the findings of our previous studies of vocabulary retention. As in the previous studies, subjects learned a list of Spanish vocabulary items using the keyword method and were assessed on three different retrieval tasks: the full retrieval task (given the Spanish word retrieve the English equivalent); the keyword retrieval task (given the Spanish word retrieve the similar-sounding English keyword); and the image retrieval task (given the keyword retrieve the English equivalent). In the current study, 24 subjects were tested, with either a one-week or one-month delay period before retest. The study was designed to assess the effects of retrieval task order (i.e., the order of the previously-mentioned full, keyword, and image retrieval tasks) on retention so that in future experiments appropriately counterbalanced sets of task orders can be selected. In addition, verbal reports were used on only half the items to assess the effects of verbalization on retention. We anticipated that verbal report items would show improved retention relative to silent items, but that the pattern of retention results for the three retrieval tasks would be

the same as in previous studies and the same for the silent and verbal items—that is the full retrieval task would be slower than either of the component retrieval tasks, and the keyword component would show better retention than the image component.

The results of our initial analyses of the data suggest the following: As anticipated, the verbal reports improved retention overall but did not significantly affect the pattern of the retention results. Also, in terms of priming, as expected, retrieval is at first mediated by the keyword and the image components. Furthermore, as in the previous studies, our analyses suggest that the image component decays more readily than the keyword component. A new and exciting result of our analyses is that the speed of retrieval on the immediate retention test is predictive of recall at delay. Finally, we have finalized our encoding scheme for the verbal protocols and have encoded a number of the protocols. The results here are also quite interesting: Reported mediation seems reliably related to speed of retrieval.

We reported our work on the keyword method at the American Educational Research Association Conference in April in New Orleans. This paper is appended to the present report (see Appendix F). The title of the paper was "A componential analysis of the keyword method;" the coauthors were Robert J. Crutcher and K. Anders Ericsson.

We have also completed the first of two new follow-up experiments, in which we are looking at what happens to the mediators as subjects become more practiced on the full retrieval task. So far, it appears that as subjects practice the full retrieval task, retrieval becomes direct (i.e., no accessing of keyword and image). We are now in the process of completing the data analysis of this study.

Having developed the above methodology for studying vocabulary retention, we are now designing an experiment to look at retention of CPR using a similar approach. In this first experiment, subjects will learn the component steps in the CPR procedure, thinking-aloud as they do so. Subsequently, they will be shown one of the component steps and asked to recall the next step. Retreival times and retrospective verbal reports will be collected to assess how one step mediates retrieval of the next step. The retrieval times should index how connected one step is to another, so that we may be able to predict delayed retention on the basis of initial retrieval time. This would provide a means to assess at inital testing whether training procedures will be effective after longer delays. In addition, the retrospective reports will enable us to determine what information from previous steps or from the overall task is being used to retrieve a step.

APPENDIX A

OPTIMIZING THE LONG-TERM RETENTION OF KNOWLEDGE AND SKILL

As those of you who were here last year know, the three of us, Anders Ericsson, Alice Healy, and I are looking for techniques to optimize the long term retention of skilled and knowledgeable performance. This is clearly not a new problem in psychology but neither is it a problem that has been solved to everyone's satisfaction. In fact, we think it's a problem that often gets overlooked by researchers and by developers of training programs who seem more commonly to focus on optimal training procedures--that is, training procedures which maximize performance in a minimal period of time--without concern for the durability of what has been learned. Thus, our purpose is in general to try to identify those conditions of training that are associated with skill or knowledge permanence and availability.

Our starting point is the observation that a significant portion of almost any learned skill or knowledge is by its very nature relatively permanent. Harry Bahrick has coined the useful descriptive term "permastore" for that portion of acquired knowledge or skill that is retained, relatively undiminished, over years of disuse or nonrehearsal. This observation presents the challenging possibility of identifying conditions of learning and/or characteristics of material or skills learned that contribute to the durability of acquired behaviors. The idea is that, if we can identify conditions or characteristics that distinguish between short-lived and relatively permanent components, we might be able to trace back and find out what aspects of training differentiate those components from other less permanent components.

Our approach to the problem is twofold. The first we call analytic. In this part of the project we started out by examining the training provided for military electronics technicians and identifying some of the interesting components of that training. We then devised laboratory tasks which seemed to capture those components. As you will see, some of the components are essentially perceptual, such as the detection of error signals on an oscilloscope. Some are largely motor, having to do, for example, with

the skillful application of test instruments to possible malfunctioning components of equipment and some of them are cognitive, involving, among other things, problem solving strategies or decisions among potential tests of malfunction. Studies undertaken on these tasks involve variations in training conditions and subsequent measures of what is retained over the long haul. We are primarily concerned with what gets retained, which is followed by a post-hoc analysis of those conditions of training that distinguish between what is retained and what is not.

The second approach we call structural. It involves an analysis of the mental structures involved in complex, natural skills and the further development of methods to characterize those structures.

Under the analytic approach, our initial goal was to devise four laboratory analogues of component skills of electronic technicians. We have in fact developed five laboratory methodologies for this purpose and have either completed or initiated the preliminary testing of each of these methodologies. The laboratory tasks involve target detection, data entry using the keypad of a typical computer keyboard, learning the logical rules involved in computing circuits, memory for numerical calculations and the temporal, spatial, and item components of memory for response sequences (see Slide 1). Although we've made substantial progress on each of these tasks, for present purposes we intend to concentrate on just one in this presentation, namely the data entry task.

We are interested here in measuring memory for the skill of entering number sequences on a keypad and any associated memory for the items entered, themselves. In the first major study, we trained 36 subjects, each of whom participated in three training sessions on successive days and a subsequent retention test one month later. They learned to type three-digit sequences, presented in blocks of 10 sequences each. Subjects were divided into three groups, depending on the extent and pattern of repeated digit sequences during training. For a control group, no

LABORATORY SKILLS

(a) Target detection

(b) Data entry

(c) Learning logical rules involved in computing circuit design

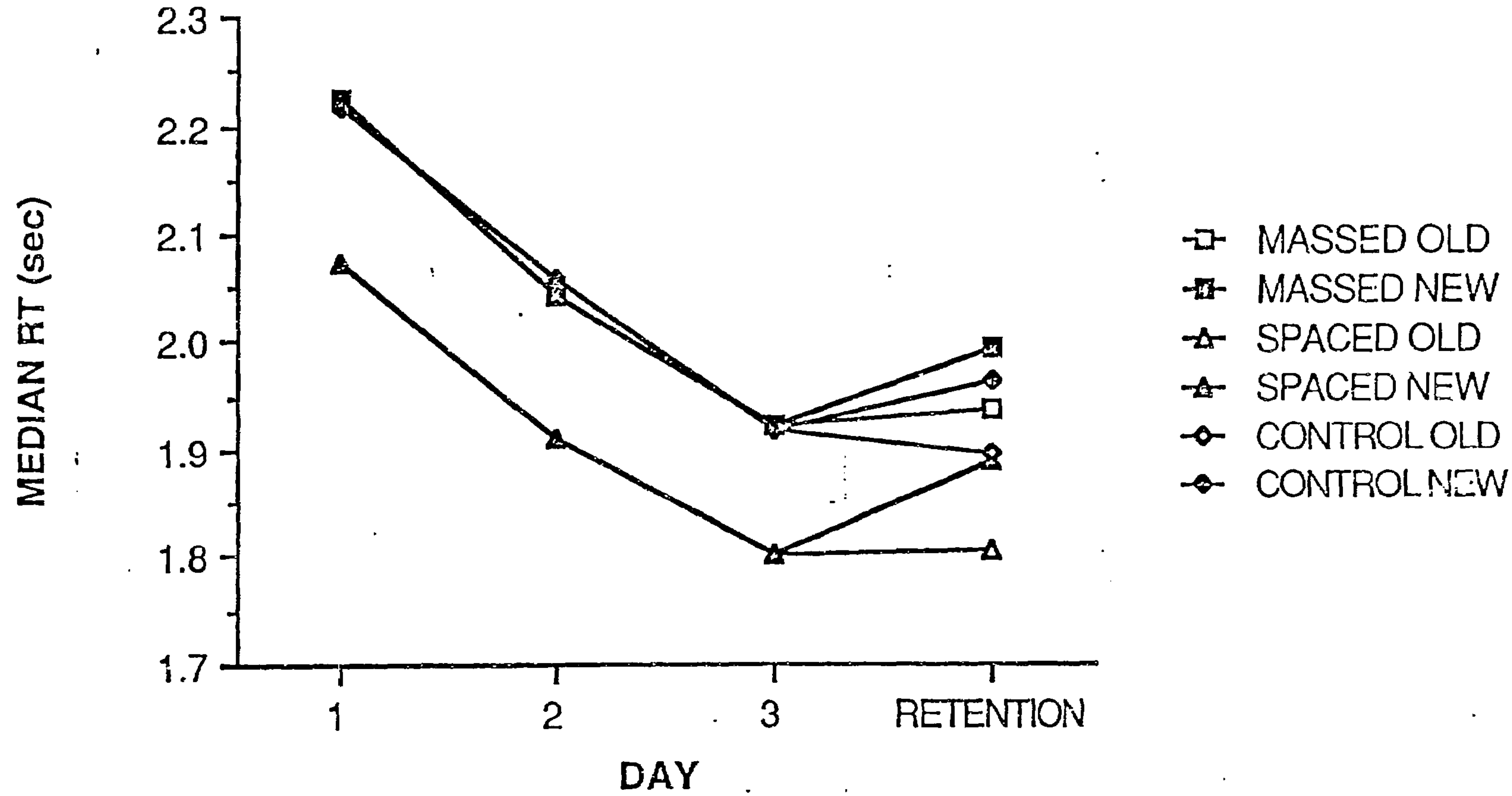
(d) Memory for numerical calculations

(e) Temporal, spatial, and item components of memory for lists

sequences were repeated. For a massed group, each of five blocks of 10 three-digit sequences was repeated five times in a row. For a space group, each block was presented five times (as in the massed condition), but, in this case, with the other four blocks intervening between repetitions. During the final retention test, subjects were given a mixture of blocks with digit sequences from acquisition trials and other blocks with new sequences not given earlier. A variety of theoretical arguments can be developed with respect to the anticipated effects of massed versus spaced repetitions. For example, an intratask interference principle suggests that subjects in the massed group should show superior performance during acquisition, but inferior performance at retention. Rather than worry about these possibilities at this point, I want to turn directly to the data. In fact, we found little difference between the three groups of subjects in terms of their overall time to enter digits during the acquisition phase (note that errors were virtually nonexistent in these data). The effects we do obtain lie in the retention test, one month later. Here, reaction times were significantly faster for the old blocks of digit sequences relative to the new blocks that had not been seen previously during acquisition (see Slide 2). This is equally true for all three groups, including the control to whom each block during acquisition was shown only once. In other words, the main effect here is a difference in responding to blocks of digit sequences that have previously been entered one or five times, relative to new sequences. Something about the acquisition sequence clearly carries over a one month period to significantly prime or facilitate later performance.

We have completed two additional experiments to follow up on this observation. In the first, the aim was primarily to replicate the effect in Experiment 1. Twenty four subjects were tested in two sessions with a one day interval between them. In each session, subjects were shown 30 blocks of 10 three-digit numbers which they were required to type on a keypad as quickly and accurately as possible. Half the blocks shown in the second session were old (i.e., they had been shown during the first

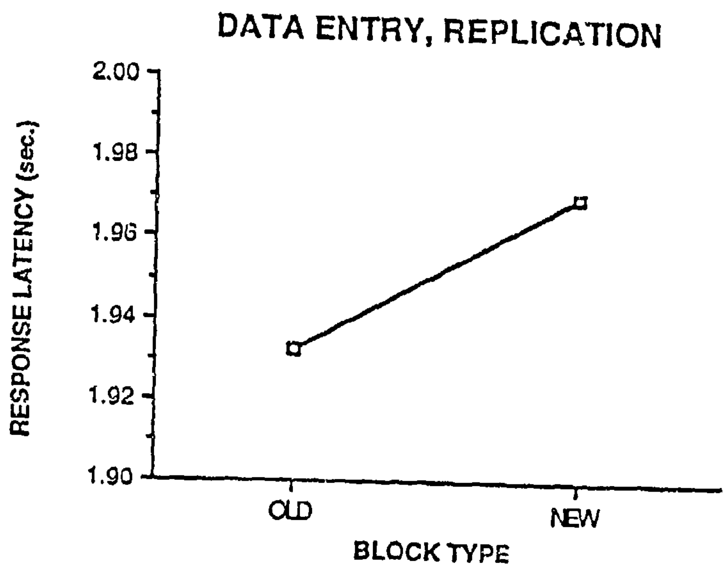
MEDIAN RT AS A FUNCTION OF GROUP AND DAY

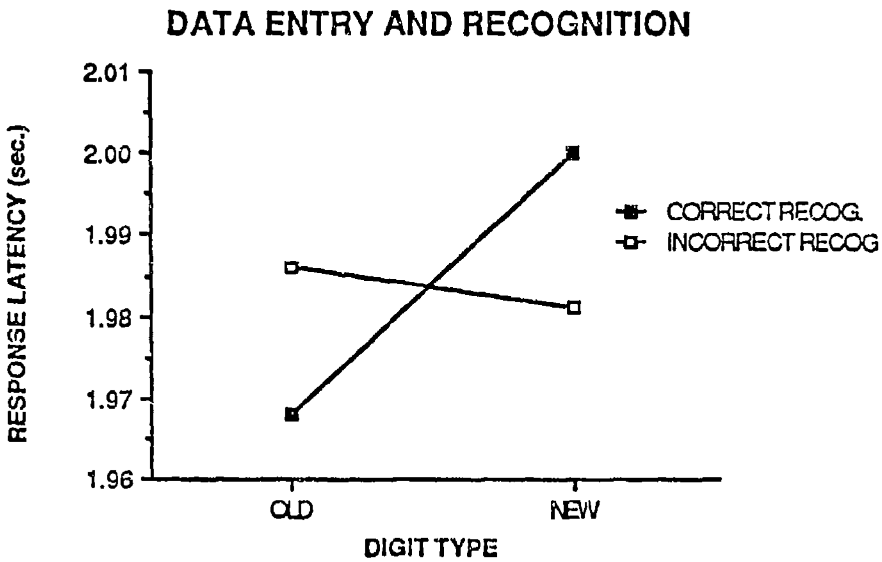


session) and half were new. Overall, typing speed was greater on the second day of training, as one would expect. Beyond that, there was a significant advantage for old blocks relative to new (see Slide 3), replicating the major observation in our initial study.

In a second follow-up study, 24 additional subjects were tested in the same manner except that the three-digit numbers were shown one at a time rather than in ten sequence blocks. In this study we were concerned with whether or not the priming effects observed in earlier studies were in any way mediated by active or conscious memory for sequences typed in the first session. Thus, on the second day of this study, subjects were asked to make old/new recognition decisions for each of the three-digit numbers immediately after typing the number. Subjects did show a low but significant level of recognition for old three-digit numbers. More importantly, however, subjects responded significantly more rapidly to old numbers only if they recognized those numbers as old (see Slide 4).

There are some fairly obvious methodological issues raised by these studies and also some fairly obvious additional questions, especially about the long term retention of the skills. But, with time limitations in mind, let me simply note at this point that we have designed three new follow-up experiments for the data entry procedure and have begun testing subjects in two of these three experiments. In the first of these, we will further explore differences between massed and spaced repetitions of stimulus items and will assess retention after a one month interval, both in terms of enhanced data entry speed and by means of explicit recognition ratings by the subjects. By obtaining recognition data sometimes before data entry, sometimes after data entry, this experiment will allow us to test the hypotheses that (a) explicit memory mediates the priming effect versus (b) apparent recognition is mediated by the subjects repetition of prior motor movements. The second experiment makes use of two different configurations of a keypad (one like that found on a typical computer console


DATA ENTRY, REPLICATION
2.00
1.98
1.96
1.94
1.92
1.90
RESPONSE LATENCY (sec.)
OLD
NEW
BLOCK TYPE


DATA ENTRY AND RECOGNITION
RESPONSE LATENCY (sec.)
2.01
2.00
1.99
1.98
1.97
1.96
OLD
NEW
DIGIT TYPE
CORRECT RECOG.
INCORRECT RECOG

and one like that found on a touch tone telephone, see Slide 5). We use this manipulation to determine if subjects retain information about the motor sequence of key presses or information about the actual sequence of digits displayed. In addition, this experiment will allow us to assess the difficulty of transferring the data entry skill from one keypad configuration to another, again after a one month retention interval. A third experiment will also employ two different key configurations, in this case the computer keypad and the less well-structured horizontal linear array of digits at the top of a standard keyboard (see Slide 6). Like the first experiment this study will examine long term memory for digit sequences both in terms of explicit recognition responses and in terms of the speed and accuracy of entering the digits. We also include in this experiment a condition in which, during training, subjects read each sequence rather than type it, pressing the space bar simultaneously with each digit. The question is whether the motor response of digit entry is necessary for facilitated recognition and/or entry responses during the retention test.

Complementary to this laboratory work, we employed the same methodology, basically, to study the retention of data entry skill learned in a natural setting. Specifically, we are working with one individual who has made extensive use of the keypad for data entry in her job of entering student's social security numbers into the computer (1 semester, ~ 10,000 entries, ~ 100 hrs of entry practice). At the conclusion of her job and then again six months later, we tested her proficiency on our task and we intend to test her repeatedly at various future dates. Our initial findings indicate no loss in speed of data entry and in fact a significant increase in accuracy over the first six month retention interval. After an additional eight month retention interval we retested the subject again. Relative to the first retention interval we found no change in accuracy (which was close to perfect) but a significant improvement in speed of responding (see Slide 7). Hence, there is no clear evidence of forgetting of this skill over fairly substantial retention intervals. A more detailed breakdown of response

KEYPAD			PHONE		
7	8	9	1	2	3
4	5	6	4	5	6
1	2	3	7	8	9
0			0		

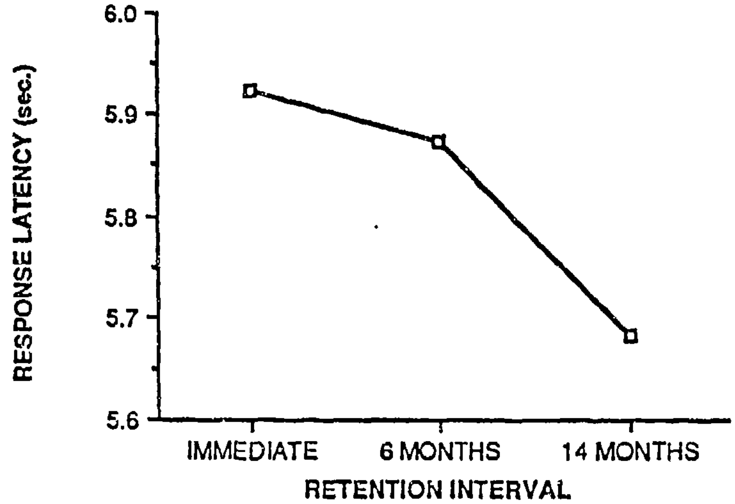
KEYPAD

7	8	9
4	5	6
1	2	3
0		

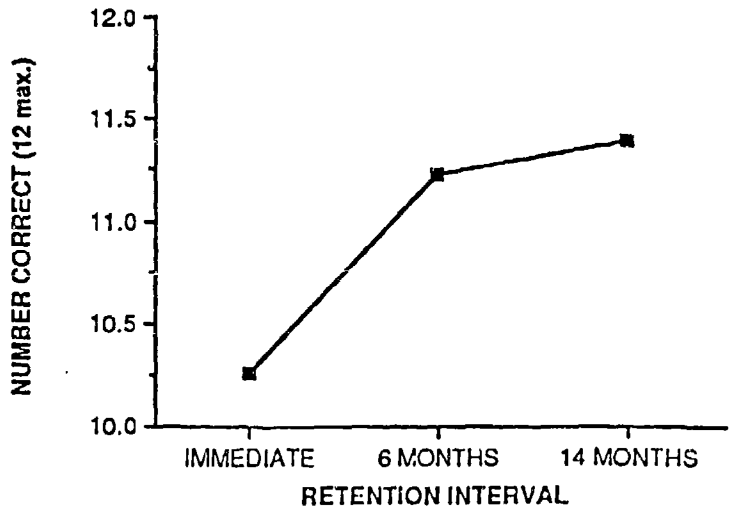
KEYBOARD ROW

1 2 3 4 5 6 7 8 9 0

S.D., DATA ENTRY

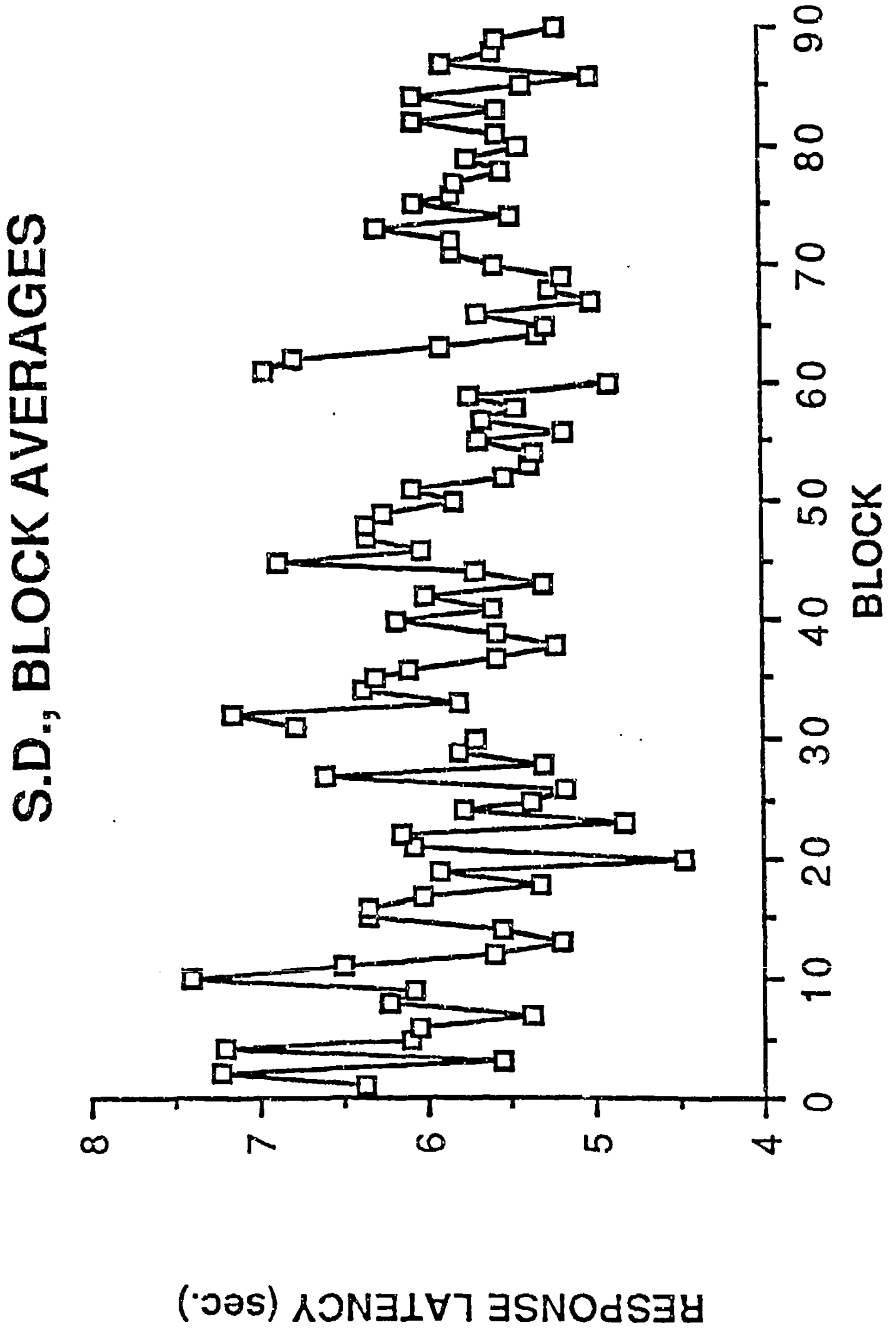


S.D., DATA ENTRY



latency as a function of block of digits revealed some initial forgetting on the retention tests, followed by rapid improvement (relearning) across the 30 blocks of a session (see Slide 8). Thus these preliminary findings for data entry under natural conditions agree with our findings for the same skill learned in the laboratory. In both cases, retention is essentially perfect and response latencies don't change over time. We hopo to find a group of students who have had similar job experiences with keyboard data entry, perhaps as cashiers or as store checkers. We intend to investigate the effects of changing context and changing tasks as well as retention intervals on the speed and accuracy of naturally acquired data entry skill.

Data entry can be studied either as a laboratory task or as a more naturally acquired skill. We have tried to identify other natural skills that have relevance to the tasks of an electronics technician and to some of the laboratory tasks we're examining. At the moment, we are concentrating on four of these natural skills. In addition to data entry, our work investigates (1) mental multiplication, (2) algebra, as it is taught in a freshman course at Colorado, and (3) the temporal, spatial, and item components of memory, in our case autobiographical memory for schedules of previously taken college classes (see Slide 9). To stay within limitations, I'll focus on our work with mental multiplication. The methodology here comes in two parts. One part uses a questionnaire, similar to that developed by Bahrick to assess when, how, and to what extent a student was originally trained on the skill of mental multiplication. The questionnaire also assesses the type and amount of maintenance activity that skill received after initial training and the variability of contexts in which the skill was maintained. The second methodological component involves an assessment of students speed and accuracy at performing simple multiplication problems in their heads. We attempt to predict performance on the assessment task from various indices of skill acquisition and maintenance derived from the questionnaire. Furthermore, following the approach we've used in laboratory tasks, we attempt to


S.D., BLOCK AVERAGES
RESPONSE LATENCY (sec.)
8
7
6
5
4
0
10
20
30
40
50
60
70
80
90
BLOCK

NATURAL SKILLS

(a) Mental multiplication

(b) Algebra

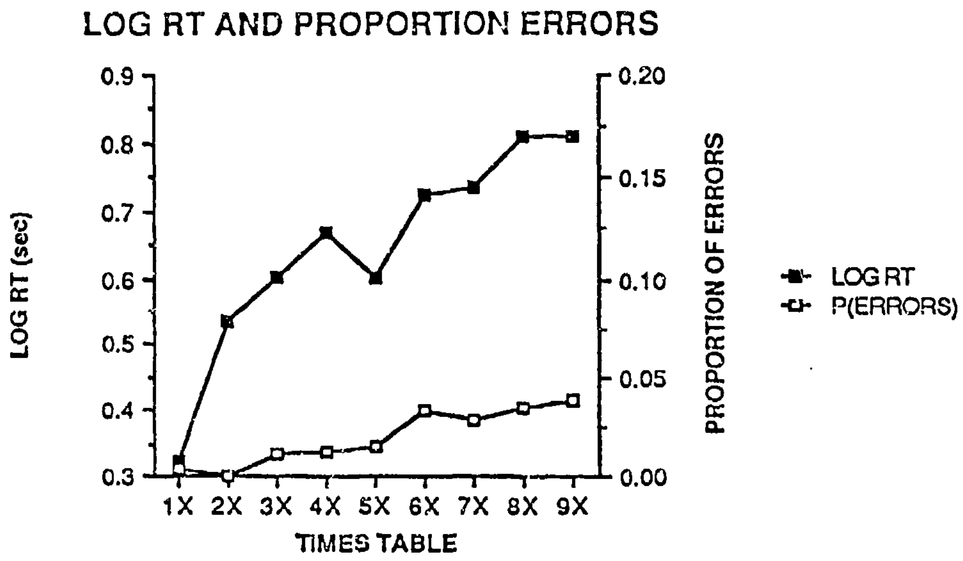
(c) Data entry

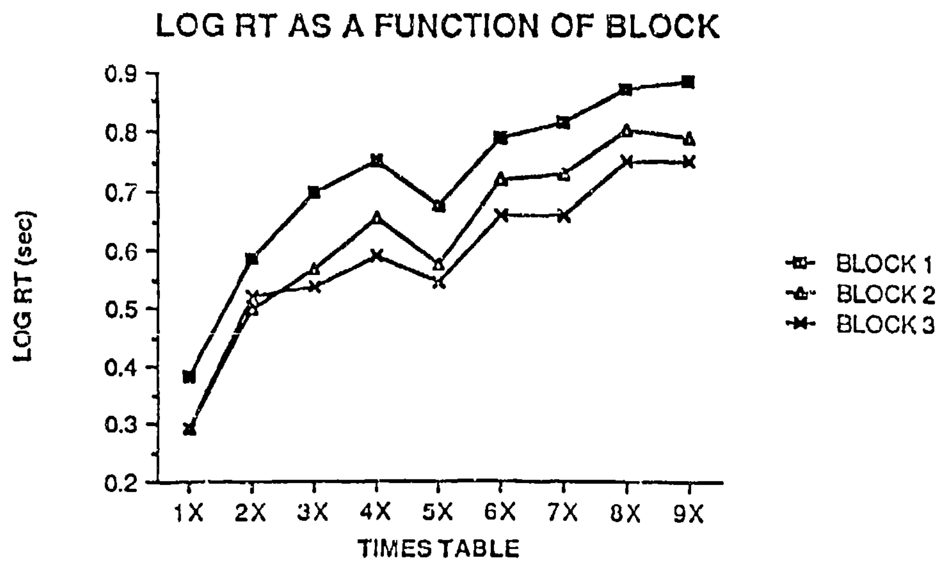
(d) Temporal, spatial, and item components of memory for class schedules

determine whether automaticity is related to skill performance and to long term skill retention. More specifically, the the first study, we systematically varied multiplication problem difficulty and used as a tentative index of automaticity the slope of the function denoting the relationship between accuracy or speed of solution and problem difficulty.

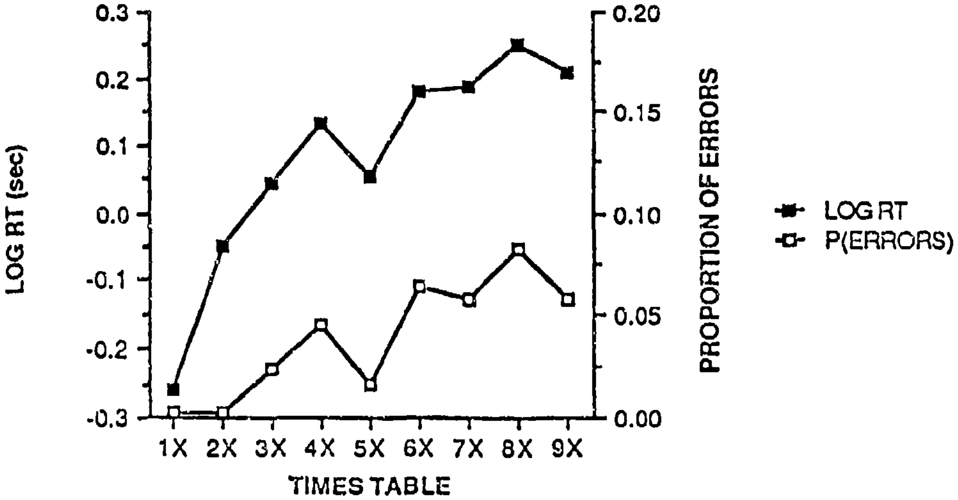
In preliminary work, we tested two groups of 9 subjects who were shown three repetitions of a complete set of single digit multiplication problems, one problem at a time on a computer terminal. Subjects in one group responded by typing the answer on a keypad, as in the data entry task. Subjects in the second group responded orally and their reaction times were measured by means of a voice key. These initial data allowed us to assess differences in problem difficulty and to examine improvements in speed and accuracy as a function of practice. In Slide 10 (keypad group) and Slide 11 (voice key group), the overall mean log reaction times and error rates (top panel) and the mean log reaction times as a function of practice block (bottom panel) are displayed.

It's clear from this preliminary study that none of our subjects achieved anything resembling the criterion of automaticity. Thus, we have initiated a long term training study with a single subject, hoping to show a progression toward automaticity with practice. Beyond that, we will assess whether the speed, accuracy and degree of automaticity are retained over long retention intervals. We've completed the acquisition phase of the study and the results are highly revealing. The subject was given 11 acquisition sessions with the keypad method of responding and then the final session using the oral method. The subject improved dramatically in multiplication skill over sessions, primarily in terms of speed of responding. Accuracy was near perfect from the outset. Differences in speed of responding as a function of problem difficulty decreased as training progressed, though they are still evident at the end of training suggesting that full automaticity was not achieved (see slide 12). We have

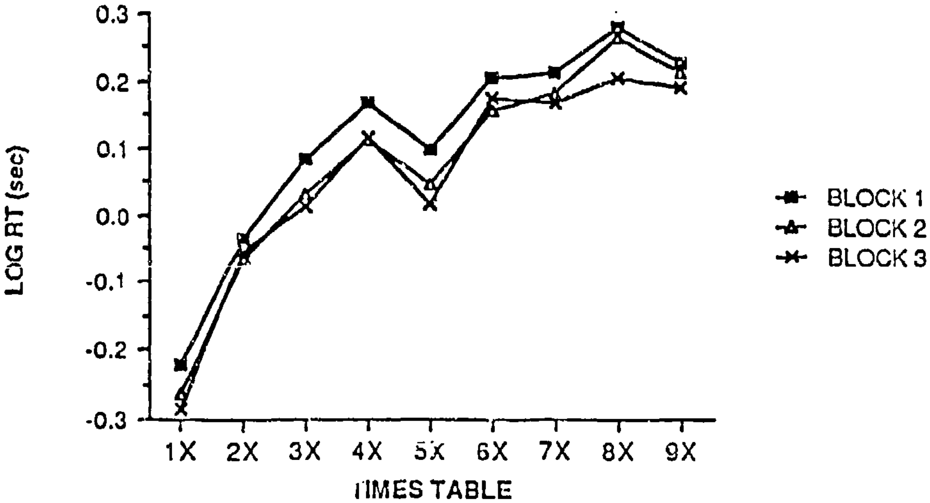

LOG RT AND PROPORTION ERRORS
LOG RT (sec)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.20
0.15
0.10
0.05
0.00
PROPORTION OF ERRORS
1X 2X 3X 4X 5X 6X 7X 8X 9X
TIMES TABLE
LOG RT
P(ERRORS)

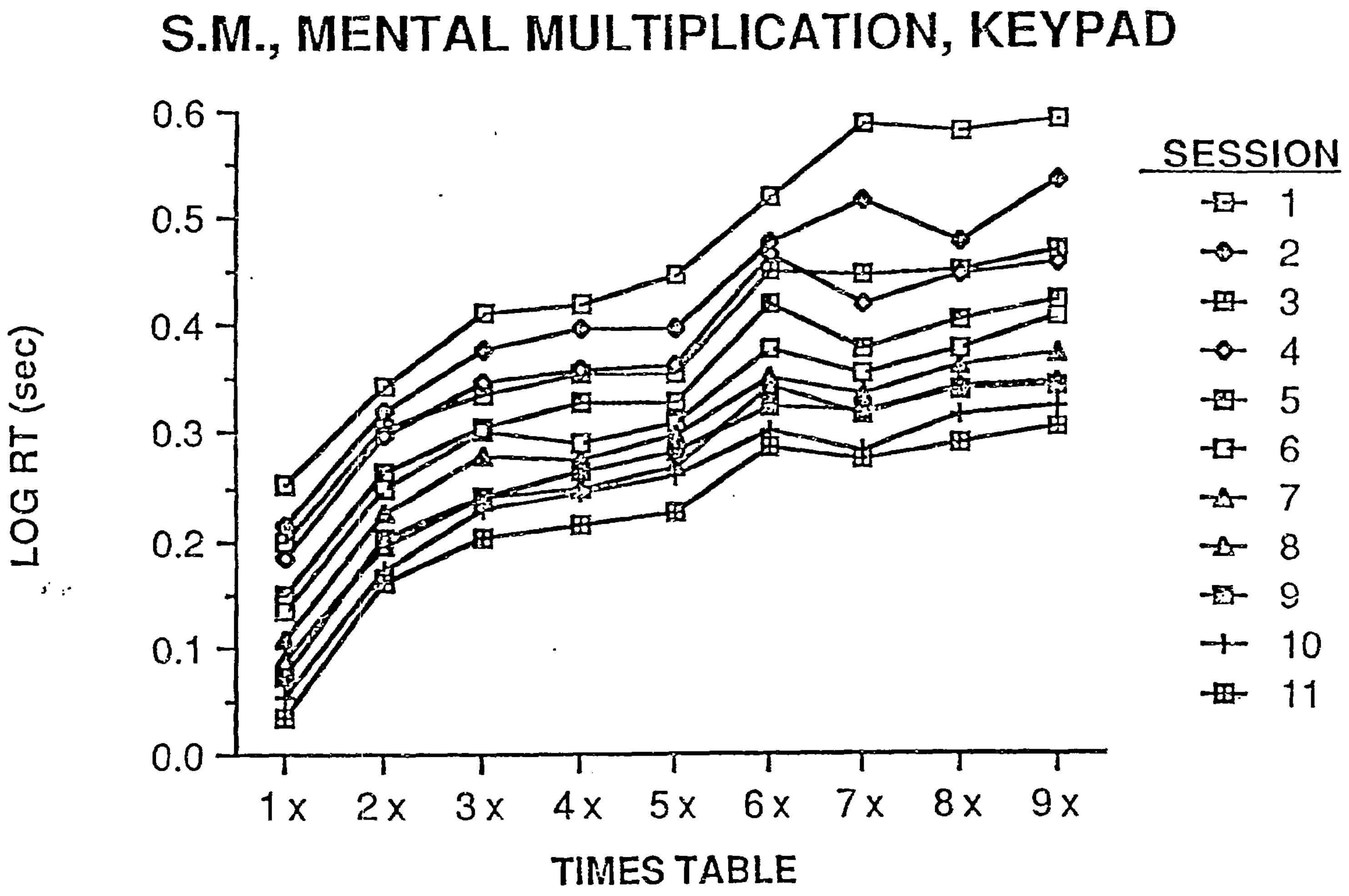

LOG RT AS A FUNCTION OF BLOCK
LOG RT (sec)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1X 2X 3X 4X 5X 6X 7X 8X 9X
TIMES TABLE
BLOCK 1
BLOCK 2
BLOCK 3

LOG RT AND PROPORTION OF ERRORS



LOG RT AS A FUNCTION OF BLOCK

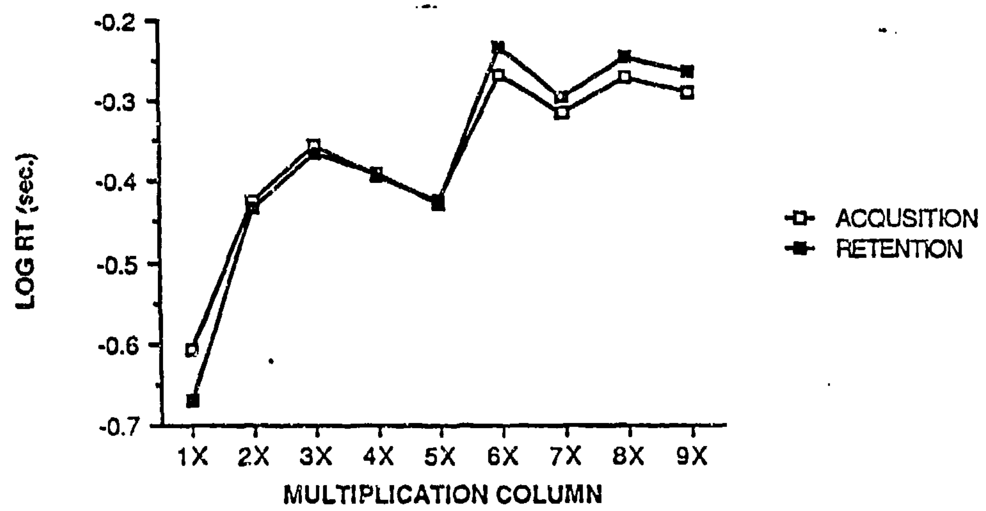



S.M., MENTAL MULTIPLICATION, KEYPAD
LOG RT (sec)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1 x
2 x
3 x
4 x
5 x
6 x
7 x
8 x
9 x
TIMES TABLE
SESSION
1
2
3
4
5
6
7
8
9
10
11

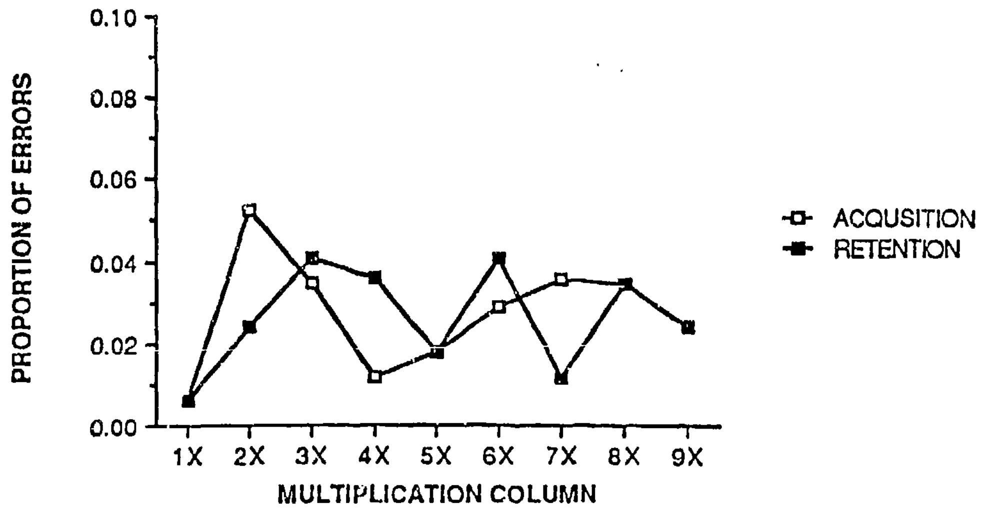
retested the subject after a 3-month retention interval using the oral method of responding. The results of this retention test are compared to the final acquisition session on slide 13. It is clear from this slide that the subject showed essentially no forgetting of this skill in terms of both speed and accuracy. Note that errors are quite low in both cases, and there is a remarkable match between the pattern of response times for the two tests. This result of perfect retention is similar to that described earlier for data entry and also like that which we discussed last year for the skill of target detection.

To test the hypothesis that long term retention of multiplication skill is influenced by the extent to which the subject has become automatic, we have designed a large follow on multi-subject experiment. Subjects will be given 10 sessions of training. Training will adopt a "drop-out procedure" designed to provide selectively more training in each session for more difficult items. In training then, we hope to force something like a flat function (or automaticity) over problem difficulty. The 10th and final session will use a normal non-drop-out procedure designed to assess the effectiveness of automaticity training for each individual. Retention of skill will be assessed 3 months later. On these data we intend to perform an individual differences analysis to determine the extent to which our measure of automaticity achieved by the last training session predicts retention. This study will also determine whether training on the task is influenced by the order of multipliers given to the subject. Specifically, some subjects will see each problem in only one of two possible orders (larger or smaller multiplier first) during the acquisition phase of training, but will have to respond to problems in both orders on the retention test. We are interested here in whether one order induces a particular strategy which then might or might not transfer to the other. The question is whether a "simplified" representation of skill generalizes to "full skill" performance.

S.M., MENTAL MULTIPLICATION, ORAL RESPONDING



S.M., MENTAL MULTIPLICATION, ORAL RESPONDING



APPENDIX B

THE ACQUISITION AND RETENTION OF A LETTER-DETECTION SKILL

Alice F. Healy and David W. Fendrich
University of Colorado

Janet D. Proctor
Purdue University

Abstract

In two experiments, we examined the acquisition and retention of a letter-detection skill with a consistent-mapping procedure. In Experiment 1, subjects were trained from 0 to 4 sessions at detecting the letter H in displays containing random letters, and retesting occurred after a one-month delay. Performance improved and in some cases became more automatic, and the performance level was maintained over the retention interval. When tested with a prose passage, the high error rate on the word THE was eliminated after training and after the retention interval, regardless of the amount of training. In Experiment 2, two subjects were given 12 sessions of training followed by a retention test 6 months later. For one subject there was also a retention test 15 months after acquisition. Performance improved dramatically with training and substantial but not complete automaticity was achieved. Performance on the retention tests was close to the final acquisition level. The surprising lack of forgetting in this study was contrasted to the substantial forgetting typically found in studies of verbal learning.

In this investigation we are concerned with the acquisition and retention of a letter detection skill. In previous research letter detection performance has been studied in two different contexts, one involving prose passages (e.g., Healy, 1976) and the other involving random letter displays (e.g., Schneider & Shiffrin, 1977). Although there has been a thorough investigation of the effects of training with random letters (including explorations of the development of automaticity), there has been essentially no research examining how training in that context affects subsequent performance in the prose context. Also, there is little known about the durability of the effects of training in the letter detection task (but see Rabbitt, Cumming, & Vyas, 1979). The present study examines the durability of the effects of training on letter detection, whether retention of the letter-detection skill depends on the amount of training or the achievement of automaticity, and the extent to which training in random letter displays influences detection in the prose context.

Dramatic forgetting is ubiquitous in verbal learning (see, e.g., Crowder, 1976), but forgetting seems to be considerably smaller in motor learning (see, e.g., McGeoch, 1942, and Naylor & Briggs, 1961; but also see, e.g., McGeoch & Melton, 1929), and relatively small in other studies of perceptual learning (e.g., Kolers, 1976). Perhaps, the learning resulting from detection training will be well retained, like motor and other perceptual learning. If so, the changes in detection performance resulting from detection training should be evident even after a relatively long delay without practice. On the other hand, if forgetting of the letter-detection skill is rapid, like verbal learning, the changes in detection performance may be transient and disappear after a delay.

Two types of letter processing have been distinguished in the literature: controlled processing, which requires attentional resources and cognitive effort, and automatic processing, which requires only minimal cognitive capacity and attention (e.g., LaBerge & Samuels, 1974; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Perhaps it will be necessary to train letter-level processing to the point of automaticity (so that letter information would be accessible without attentional resources) in order for superior long-term retention of the letter-detection skill. Alternatively, the amount of forgetting may not depend on whether automatic processing develops.

In previous studies of letter detection in prose, two striking findings have been well documented. First, a "word frequency disadvantage" has been found, in which letters occurring in very common words (such as <u>the</u>) are missed more often than those occurring in less common words (such as <u>thy</u>) (e.g., Healy, 1976). Second, a "word inferiority effect" has been found, in which letters are more likely to be missed in correctly spelled words (again, such as <u>the</u>) than in misspelled words (such as <u>teh</u>) (e.g., Healy & Drewnowski, 1983). Underlying the explanations of these effects (see, e.g., Drewnowksi & Healy, 1977; Healy & Drewnowski, 1983; Healy, Conboy, & Drewnowski, 1987; McClelland & Rumelhart, 1981) is the basic assumption that the failure to detect letters in common, correctly spelled words results from interactions of processing at the word and letter levels. This assumption not only predicts that enhancing processing at the word level may inhibit further processing at the letter level, but also leads to the prediction that enhancing processing at the letter level will change the pattern of letter-detection errors. Indeed, in previous research with the letter detection task (Healy, Oliver, & McNamara, 1987), the pattern of errors has been found to be changed as a result of practice. Specifically, a decrease in the overall error rate and in the size of both the word frequency disadvantage and the word inferiority effect was found as a function of repeated

exposure to the same prose passage. However, this effect of practice might have been the result of familiarity with the specific passage rather than the result of improved letter-level processing, especially because the effects of passage familiarization have been found to be substantial in studies of proofreading for misspellings (see Levy, 1983; Levy & Begin, 1984; Levy, Newell, Snyder, & Timmins, 1986). Hence, it is important to construct a situation in which only letter processing is practiced, so that the effects of training at the letter level can be assessed.

The aim of the present study was to examine these issues concerning the acquisition and retention of a letter-detection skill by constructing a task analogous to that used with letter detection in prose but in which only letter processing was practiced. We achieved this end by developing a variant of the detection training paradigm developed by Schneider and Shiffrin (1977). Specifically, as in the prose letter-detection task, character sequences were rapidly presented on a computer terminal screen, and subjects pressed a response key when they detected the target letter (see, e.g., Healy, Oliver, & McNamara, 1987; Proctor & Healy, 1985). In this case, however, random letter sequences, rather than connected text, were employed. Further, as in the detection training paradigm, frame size (the number of letters in each display) was varied, yielding slower and less accurate responding with larger frame sizes. Automatic processing was indexed by a decrease in the effect of frame size as a function of practice. Before and after detection training, subjects were exposed to the standard letter-detection task with a prose passage. We expected that the word frequency disadvantage and word inferiority effect would be evident before training. However, these effects should be reduced or eliminated after training, especially if subjects became automatic at letter detection. In our experiments subjects also returned for additional testing after a long delay interval. During this retention test they were given another exposure to the

prose letter detection task as well as the random-letter task that they had practiced. We expected that the effects of training on letter detection performance in both tasks would be well maintained across the delay interval if this skill resembles other perceptual and motor skills in its retention characteristics.

Experiment 1

In preliminary research (Healy, Fendrich, & Proctor, 1987) we found that the word frequency disadvantage in the letter-detection task with prose passages was large in a pretest but was eliminated on a posttest after detection training. One purpose of Experiment 1 was to compare the effect of two different amounts of detection training. Perhaps the word frequency disadvantage will only be eliminated if the subjects are given sufficient practice so that their performance at least approaches automaticity. To address this question, we included two experimental groups of subjects who were exposed to different amounts of detection training; either two or four days of training were administered before the posttest.

The second and most important purpose of Experiment 1 was to examine the permanence of the effects of detection training. Towards this end, we employed a retention-test phase approximately one month after the posttest. The retention tests included letter detection in a prose passage, followed by the detection training task. The retention test with the detection training task allowed us to examine the durability of the detection skill across a lengthy delay interval. The retention test with the prose passage allowed us to assess the durability of the changes in the pattern of letter-detection errors resulting from detection training. Perhaps the word frequency disadvantage will be eliminated at the end of training but will reappear in the retention test. Alternatively, the skill learned during detection training may be retained so well that the pattern of results on the posttest will persist to such an extent

that the word frequency disadvantage will continue to be absent during the retention test.

Method

Subjects

Thirty-six students at the University of Colorado participated for course credit in Introductory Psychology and for payment at the rate of $5.00 per hour for any additional hours beyond the course requirement.

Stimuli and Apparatus

Detection training stimuli. The detection training displays were strings of 16 letters and two internal blank spaces (see Figure 1). This length corresponded to the approximate length of the letter-detection passage displays. Each string contained 2, 4, or 16 scrambled uppercase letters, depending on the frame size (2, 4, or 16), randomly interspersed with 14, 12, or 0 filler characters which were number signs (#). A target character (*H*) was present in half of the character strings. The two blank spaces were randomly placed in each string, with the constraint that they could not occur in the first, the last, or adjacent string positions. These blanks gave the displays the appearance of the three-word configuration of the letter-detection passages. The nontarget letters (distractors) used were the same as those used in the second letter-detection passage except that they were scrambled at random.

Insert Figure 1 about here

Letter-detection passages. Three prose passages in uppercase type were employed. One passage was adapted from a passage of Winston Smith's novel, *The Stranger from the Sea*. The text contained 483 words, including 72 test words containing the target letter *H*. The word *THE* accounted for 36 of the test words. The remaining 36 test words were other lower frequency words containing

FRAME SIZE	
2	## ####### #####HI#
4	##O ## H####I##M###
16	WYSEYIG PEO PCNUHE

Figure 1. Example from experiment 1

a single H. At most one test word occurred in each three-word display segment, and test words were located with equal frequency in all three positions.

Half of the test words of each type were misspelled, and two versions of the passage were produced by varying which half of the words were misspelled. As a result, a word-nonword comparison and an examination of the effects of word familiarity were made possible without the confounding variables of word length and frequency of occurrence in the text. Twelve nontarget, filler words also were misspelled so that incorrect spelling would not automatically signify the presence of a target. The same filler words were misspelled in both versions of the passage, according to a prescribed procedure. The last letter of the word was replaced with another letter, unless the last letter was a target. In that case, the first letter was replaced. Original letters and substitutes were paired so that the same substitution was always made for a given letter (e.g., THE was always misspelled THD), except when that substitution would produce another word. In those cases, an alternative substitute letter was selected.

The second passage was adapted from another portion of The Stranger from the Sea. It contained 783 words, including 48 occurrences of the word THE and 90 other lower frequency words containing a single H. Test words occurred with equal frequency in all three positions in a segment, and at most one test word appeared in each segment. The misspelling procedure described above was implemented, except that 15 filler words (rather than 12) were misspelled.

One of the first two passages was expanded and modified to create the third passage. The third passage contained 1,296 words, including 204 test words containing the letter H. The word THE accounted for 102 of the test words. The remaining 102 test words were other lower frequency words containing a single H. Half of the test words of each type (THE and other) and 24 additional filler words were misspelled using the procedure described above. Test words of each type and spelling occurred equally often ($n = 17$) in each of the three word

positions in a display segment of text.

Reading comprehension tests were constructed for each passage. Each test contained eight moderately difficult, four-alternative multiple choice questions.

Apparatus. Except for the reading comprehension questions, which were presented on paper, all stimuli were presented on a Visual 200 cathode-ray tube (CRT) display screen linked to a PDP-11/03 computer system. The computer controlled stimulus presentation and recorded response latencies. Each subject responded by pressing a button held in his or her preferred hand. Measured from a viewing distance of 50 cm, the mean length of a line of text across all three passages was 4.96 deg of visual angle, and detection training stimuli subtended 5.27 deg of visual angle. Single uppercase letters subtended 0.23 deg horizontally and 0.46 deg vertically. A 0.34 deg space occurred between words and in detection training stimuli.

Procedure

General design. Three groups of 12 subjects each participated in two, three, or five sessions conducted over approximately three to five weeks. Group assignment was made according to a prescribed rotation based upon a subject's time of arrival. Because the experiment was conducted during two school terms (Spring, Summer), equal numbers of subjects from each term were assigned to each group in order to counterbalance any extraneous factors arising from different student populations. The three groups only differed in respect to the amount of detection training subjects received. The control group received no detection training, whereas the limited training group received two days of training (10 blocks), and the extensive training group received four days of training (24 blocks).

A standard sequence of tasks was used with all subjects, although the control group did not participate in the detection training phase. Table 1 shows the specific order and timing of tasks for each group. The experiment began with a pretraining letter-detection task using a prose passage. The first session continued with the initiation of the detection training phase. Because subjects in the control group received no detection training, they proceeded directly to the next task (posttest prose passage letter-detection task). Subjects in the limited training group performed five blocks of detection training during the first session, and an additional five blocks two days later in the second session. The extensive training group subjects received five blocks of training during both the first and fourth sessions, and seven blocks during both the second and third sessions. For the extensive training group, Sessions 1 and 2 were separated by two days, as were Sessions 3 and 4; five days separated Sessions 2 and 3.

Insert Table 1 about here

After the detection training phase was completed, subjects immediately performed a posttraining letter-detection task with a second prose passage. A retention interval of three to five weeks then elapsed before subjects returned for the final (retention) phase of the experiment. At that time, a third passage was presented for a retention test of letter detection in prose. Next, subjects (including the control group) performed five blocks of the detection training task to evaluate retention of the letter-detection skill.

Detection training. Letter strings were presented briefly in the approximate center of the terminal display screen using a variation of the rapid serial visual presentation procedure, as in the prose passage letter-detection task. Three frame sizes (2, 4, 16; number signs filled any remaining character

Table 1

Order and Timing of Tasks: Experiment 1

	Task			
Group	Pretest session	Training duration	Posttest session	Retention duration
Control	Session 1	0 sessions	Session 1	3-5 weeks
Limited	Session 1	2 sessions	Session 2	3-5 weeks
Extensive	Session 1	4 sessions	Session 4	3-5 weeks

spaces) were employed. Subjects were instructed to press a button as rapidly as possible whenever the target (H) occurred.

Each training session was organized into several blocks of trials. A training block consisted of three sets of 52 trials (26 target, 26 nontarget), one for each frame size. Frame size order was random within each block. Stimulus exposure duration was 1500 ms throughout training.

Prose passage letter-detection tasks. The text was presented three words at a time in the approximate center of the computer terminal screen, using a variation of the rapid serial visual presentation procedure (Forster, 1970). Each three-word segment was presented for 1500 ms. Within each training group, half of the subjects received one passage version, and half received the other. Subjects were instructed to read for comprehension and to press a button as rapidly as possible whenever the target letter occurred. Subjects searched for H in each passage. Passage order was counterbalanced across subjects. Reading comprehension questions were administered in a multiple-choice, paper and pencil format immediately following each passage.

Results

Scoring Procedures

Because the rapid serial visual presentation procedure has essentially no interstimulus interval, a delayed response to one stimulus can be registered during the presentation of the following stimulus. A response latency criterion was adopted to prevent including this type of response in the data. A response was considered to be correct (hit) and was included in the calculation of the response latency and accuracy data if the response was made during the presentation of a target stimulus and the response latency exceeded 200 ms. Responses made after the first 200 ms of a display presentation that did not include a target letter were scored as false alarm errors. All responses with latencies under 200 ms were not scored and were eliminated from further

analysis.

Detection Training

The proportion of hits and the median response latencies for hits were computed for each subject as a function of test block and frame size. Daily means for each training group are shown in Tables 2 and 3. The standard errors of the mean proportion of hits in Table 2 are .010 for the limited training acquisition sessions and .007 for the extensive training acquisition sessions. The standard errors of the mean response latencies in Table 3 are 10 ms for the limited training acquisition sessions and 8 ms for the extensive training acquisition sessions. The standard errors of the retention data from all three groups in Tables 2 and 3 are .015 and 13 ms, respectively. The proportion of false alarms were computed but not analyzed further due to their very low frequency (mean = .03 for the limited training acquisition sessions, .02 for the extensive training acquisition sessions, and .02 for the final retention session of all three groups). Because of problems of interpretation due to ceiling effects on accuracy with Frame Sizes 2 and 4, we present the statistical analyses of the response latency data only.

Insert Tables 2 and 3 about here

Training phase. Subjects received different amounts of detection task training depending on their condition. As a result, an overall analysis of training including all subjects could not be performed. Instead, the data from the training period for the limited (2-day) and extensive (4-day) training groups were analyzed separately initially to evaluate the development of processing automaticity.

Table 2

Mean Proportion of Hits as a Function of Training Group, Frame Size, and Day of Training: Experiment 1

		Day of Training				
Group	Frame Size	1	2	3	4	Retention
Control						
	2					.98
	4					.93
	16					.77
Limited						
	2	.99	.98			.98
	4	.94	.96			.93
	16	.64	.73			.76
Extensive						
	2	.99	1.00	.99	1.00	1.00
	4	.95	.98	.98	.98	.99
	16	.74	.82	.85	.89	.88

Table 3

Average Median Response Latency (in Milliseconds) as a Function of Training Group, Frame Size, and Day of Training: Experiment 1

Group	Frame Size	Day of Training 1	2	3	4	Retention
Control						
	2					676
	4					781
	16					1003
Limited						
	2	660	622			624
	4	770	728			729
	16	1017	992			992
Extensive						
	2	687	631	628	600	633
	4	805	721	711	685	718
	16	1000	959	931	915	903

For both the limited and extensive training groups, the hit rate was highest with small frame sizes and increased as training progressed. The effect of frame size diminished with training, and thus some progression toward automaticity did occur. This shift toward automaticity was minimal, however; the magnitude of the frame size effect was reduced only slightly, and a substantial difference remained between Frame Size 4 and Frame Size 16 when training ended.

As opposed to the accuracy measure, the response latency data of the limited training group gave no evidence for improved automaticity in a 2 X 3 (Day of Training X Frame Size) analysis of variance. The main effects of day of training, $F(1, 11) = 10.31$, $p < .01$, and frame size, $F(2, 22) = 292.17$, $p < .001$, were significant, but the Day of Training X Frame Size interaction did not even approach significance.

A similar pattern was present for the extensive training group in a 4 X 3 (Day of Training X Frame Size) analysis of variance. Response latencies decreased with training, $F(3, 33) = 17.44$, $p < .001$, and responses were slower to larger frame sizes, $F(2, 22) = 219.33$, $p < .001$, but the Day of Training X Frame Size interaction was not significant.

Retention phase. To evaluate the extent to which the effects of detection training were retained over time, the limited training and extensive training groups' performances from the last detection training session were compared to those from the retention session. These data are included in Tables 2 and 3. The previous separate analyses of detection training did not allow a direct comparison of the degree to which automaticity was attained by the two groups. The current retention analyses do provide this comparison and indicate that, although full automaticity was not achieved, the extensive training group reached a significantly greater degree of automaticity than did the limited training group.

Subjects who received extensive training had higher hit rates and smaller frame size effects on hit rates (greater automaticity) than subjects who received limited training. There was no difference between the hit rates on the last training session and the retention test, reflecting both groups' almost complete retention of letter-detection skills in the detection training task over three to five weeks. Further, the level of automaticity did not change over the retention period for either group.

Response latencies for the last training session and the retention test were compared in a 2 X 2 X 3 (Training Group X Day X Frame Size) analysis of variance. This analysis yielded only two significant effects, the main effect of frame size, $\underline{F}(2, 44) = 488.29$, $\underline{p} < .001$, and the Training Group X Frame Size interaction, $\underline{F}(2, 44) = 6.58$ $\underline{p} < .01$. Response latencies increased as frame size increased, and this effect was larger for subjects given limited training than for those given extensive training, suggesting more automatic responding for the subjects exposed to more training. This interpretation is supported by a trend analysis which revealed a significant Training Group X linear Frame Size interaction component, $\underline{F}(1, 22) = 8.11$, $\underline{p} < .01$.

Although subjects in the control group received no detection training during the main training sessions, they did receive five blocks of training during the retention test session. These data were used in a final set of analyses to compare the control group's performance during one session of detection training with the performance after a retention period for groups that had received limited or extensive training. To provide a more detailed account of any changes in performance, the data from each block of trials were used, rather than the daily averages used previously. These data are shown in Tables 4 and 5. The standard error of the mean proportion of hits in Table 4 is .017 and of the mean response latencies in Table 5 is 19 ms, as determined by analyses of variance.

Insert Tables 4 and 5 about here

The hit rate of the control group improved over the session, whereas the performance of the limited training group decreased slightly. Further, the extensive training group had a higher hit rate in general and a smaller frame size effect than the control and limited training groups had. For the proportion of hits, therefore, the limited training group performed no better after the retention interval than the control group performed during its initial training, but the extensive training group performed at a higher level and was more automatic than the groups receiving less training.

The data for response latency from the overall 3 X 5 X 3 (Training Group X Block X Frame Size) analysis of variance were partitioned into comparisons of the control group versus the limited training group, and the combination of the control group and the limited training group versus the extensive training group. Only the Training Group X Block term in the comparison of the control and limited training groups was significant, $F(4, 132) = 4.13$, $p < .01$. As with the accuracy measure, the control group's performance improved across blocks, whereas the limited training group's performance worsened somewhat. When the control and limited training groups were combined and their response latencies were compared to those of the extensive training group, only the interaction of training group and frame size was significant, $F(2, 66) = 6.78$, $p < .01$. A trend analysis revealed that this interaction included a significant Training Group X linear Frame Size component, $F(1, 33) = 8.58$, $p < .01$. The main effect of training group was not significant. Although the overall level of response latency did not differ between the extensive training group and the groups receiving less training, the magnitude of the frame size effect was smaller for the extensive training group, and thus, automaticity of processing was greater.

Table 4

Mean Proportion of Hits in Retention Test as a Function of Training Group, Frame Size, and Trial Block: Experiment 1

Group	Frame Size	Trial Block 1	2	3	4	5
Control						
	2	.96	.98	.98	.97	.99
	4	.90	.91	.94	.95	.96
	16	.74	.76	.79	.76	.79
Limited						
	2	.99	1.00	.97	.97	.97
	4	.97	.95	.95	.90	.90
	16	.76	.79	.76	.74	.74
Extensive						
	2	1.00	.99	1.00	1.00	.99
	4	1.00	.98	.99	.98	.98
	16	.93	.88	.88	.84	.88

Table 5

Average Median Response Latency (in Milliseconds) for Retention Test as a Function of Training Group, Frame Size, and Trial Block: Experiment 1

Group	Frame Size	Trial Block 1	2	3	4	5
Control						
	2	699	700	650	664	665
	4	864	791	756	746	746
	16	1026	1013	1013	991	972
Limited						
	2	625	600	597	641	656
	4	703	721	733	755	733
	16	984	1006	1013	1001	958
Extensive						
	2	610	634	633	633	656
	4	709	719	708	724	732
	16	918	924	916	860	897

Letter Detection in Prose

The proportion of targets detected (hits) was computed for each subject as a function of test-word type (THE/other) and test-word spelling (correct/misspelled). As in the detection training analyses, a response latency criterion of 200 ms was adopted. For the present analyses, all latencies under 200 ms were treated as failures to respond, as in previous studies. Group means are shown in Table 6. The standard error of the mean proportion of hits in Table 6 is .023, as determined by an analysis of variance. The mean proportion of false alarm errors overall was quite small (mean = .03); hence these false alarm data will not be further discussed.

Insert Table 6 about here

Because all groups received the same three passage tests, the data from all subjects could be combined in one overall 3 X 3 X 2 X 2 (Training Group X Test X Word Type X Spelling) analysis of variance.

The main effects of word type, $F(1, 33) = 4.43$, $p < .05$, and spelling, $F(1, 33) = 85.65$, $p < .001$, were significant, with a greater proportion of hits overall on THE relative to other, less common words (a word frequency advantage), and on misspelled words than on correctly spelled words (a word inferiority effect). Most importantly, the Word Type X Spelling interaction was significant, $F(1, 33) = 75.34$, $p < .001$. As in previous experiments, subjects made the lowest proportion of hits on correctly spelled instances of the common word THE, a somewhat greater proportion of hits on other correctly spelled words, and the greatest proportion of hits on misspelled instances of THE and other words. Thus, the word inferiority effect was greater for the word THE than for other words.

Table 6

Mean Proportion of Hits as a Function of Training Group, Test, Word Type, and Spelling: Experiment 1

		Pretest		Posttest		Retention	
Group	Word	Cor	Mis	Cor	Mis	Cor	Mis
Control							
	the	.54	.81	.67	.92	.74	.91
	other	.67	.75	.68	.72	.70	.75
Limited							
	the	.39	.76	.57	.92	.60	.88
	other	.63	.70	.68	.73	.62	.68
Extensive							
	the	.47	.81	.70	.91	.70	.95
	other	.68	.73	.75	.79	.76	.81

Note. Cor = Correct; Mis = Misspelled.

The results of the pretraining letter-detection task nicely match those of previous studies (Healy, Oliver, & McNamara, 1987; Proctor & Healy, 1985). Targets were less likely to be detected when the target was in the context of a correctly spelled, very high frequency word, such as THE. This replication extends the generality of the word frequency disadvantage and word inferiority effect to new passages and a new target letter.

Performance changed as a function of test. The main effect of test day was significant, $F(2, 66) = 18.49$, $p < .001$, with proportion of hits increasing from the pretest to the posttest, and then remaining relatively unchanged in the retention test. Also, test interacted with word type, $F(2, 66) = 32.39$, $p < .001$, and with spelling, $F(2, 66) = 3.62$, $p < .05$. On the pretest, fewer targets were detected with the word THE than with other words (a word frequency disadvantage), but on the posttest and retention test the opposite pattern occurred (a word frequency advantage), due primarily to a large increase in accuracy with the word THE (correctly spelled and misspelled) but only a modest increase in accuracy with other words. This type of reversal did not occur for the effect of spelling, but the difference between the proportion of hits made on misspelled words and on correctly spelled words decreased across tests. That is, the word inferiority effect decreased in magnitude with subsequent tests, although it remained substantial.

The length of training on the detection task had no significant effect on the proportion of hits for letter detection in prose. Neither the main effect of training group, nor any of its interactions were significant.

Discussion

As in our preliminary study (Healy, Fendrich, & Proctor, 1987), the word frequency disadvantage was eliminated after detection training, and the word inferiority effect was reduced in magnitude. However, Experiment 1 provides no support for the hypothesis that the change in these effects was due to the

detection training itself. The group receiving the most extensive detection training did not perform differently on letter detection in prose than did the groups that received limited or no training. In fact, the earliest loss seemed to occur for the control group, which received no detection training prior to the presentation of the passages. This finding suggests that experience with letter detection in prose, itself, is the critical factor. Further, passage familiarity cannot be the basis for this effect because a given subject saw a different passage at each testing. Most crucially, it should be noted that the change in performance was not short lived; the word frequency disadvantage did not reappear even after a retention interval of a month.

The disappearance of the word-frequency disadvantage as a result of experience with the prose letter-detection task may at first appear to be problematic for the unitization hypotheses (see, e.g., Healy, Oliver, & McNamara, 1987), because these hypotheses were developed specifically to account for the preponderance of letter-detection errors on frequenct words. However, in fact, the findings from Experiment 1, although unexpected, do not pose a serious threat to the unitization hypotheses. According to these hypotheses, text is processed in parallel at the level of letters and words. Because of their familiar visual configuration, very common words like <u>the</u> may be identified before their component letters. Once a word unit has been identified, the processing of the component letter units is terminated even if they have not yet reached the point of identification. This premature termination leads to errors on the letter-detection task and is caused by the pull of the text resulting from the comprehension processes. In this way the unitization hypotheses can account for the preponderance of letter-detection errors on common words like <u>the</u>, the word frequency disadvantage. How can these hypotheses accommodate the loss of the word frequency disadvantage found with prior exposure to the prose letter-detection task? An explanation of this

finding can be made simply by proposing that the pull of the text can be weakened with practice at the prose task, so that subjects learn to continue processing at the letter level even when word-level processing has already been completed. In other words, the compulsion to move on in a text at the expense of letter-level processing apparently can be diminished with experience at a task requiring letter identification in the context of reading prose. Detection training outside the prose context presumably cannot affect the compulsion to move on because there is no text, and hence no pull of the text, in that situation.

Although extensive detection training did not have a greater effect on the prose task than did limited training, it did produce greater automaticity in Experiment 1. Full automaticity was not obtained, however. We were surprised by this finding because it has been said that automaticity frequently develops in about 200 consistent-mapping trials or after two hours of training (Schneider & Fisk, 1983). Our extensive training group had considerably more practice than required by these norms. Further, previous studies (e.g., Dumais, 1979; Schneider & Shiffrin, 1977, Experiment 2) obtained automaticity for response latency with processing loads (memory set size X frame size) of up to 16 characters (and with frame sizes of up to 16 characters in the study by Dumais), and the amount of training in these studies was comparable to that used in Experiment 1. One difference between the present experiments' procedure and that of Schneider and Shiffrin (1977) is the physical arrangement of the stimuli. Whereas they used a central fixation and presented the letters in a square around fixation, in the present experiment we displayed the stimuli in a string extending from left to right, a format similar to that found in normal text. In addition, the display size used by Schneider and Shiffrin allowed subjects to view all characters with high acuity in a single fixation, whereas the display size used in the present experiment presumably required several

fixations in order for all characters to be seen with high acuity (see Shiffrin & Schneider, 1977, p. 166, for a discussion of this issue).

We wondered whether the extent to which the stimulus falls in peripheral vision, the density of the letters, or some other aspect of the stimulus itself precluded the development of complete automaticity in Experiment 1. In order to test this hypothesis, we conducted a follow-up experiment (see Healy, Fendrich, & Proctor, 1987) which made use of distractor letters (O) that were maximally discriminable from the target letter (H). Specifically, this experiment included detection training like that used in Experiment 1 and the same procedure except that the distractor letters were always O. In the context of these distractor letters, unlike the random distractor letters used in Experiment 1, we predicted that the target letter would "pop out" (see, e.g., Gardner, 1973; Treisman & Patterson, 1984) and no disadvantage for Frame Size 16 would be evident even with minimal training, unless the disadvantage was due solely to stimulus display characteristics which could not be overcome. In fact, we found that the large disadvantage for Frame Size 16 was eliminated in this follow-up experiment, so that performance was no worse (indeed was better) on Frame Size 16 than on Frame Size 4. Therefore, we concluded that the frame size effect, and hence the failure to find complete automaticity, in Experiment 1 cannot be attributed to artifacts concerning visual angle and other characteristics of the visual display.

In any event, the degree of automaticity did not seem to be related to the degree of long-term retention in Experiment 1. The limited training group, like the extensive training group, showed essentially perfect skill retention, even though there was evidence of automaticity (albeit weak evidence) only for the extensive training group. In fact, as mentioned above, all three groups of subjects, including the control group who was given no detection training, showed retention of the loss of the word frequency disadvantage over the

one-month retention interval.

Experiment 2

In Experiment 1 we found essentially no forgetting of the letter-detection skill in terms of both speed and accuracy over a one-month retention interval. Hence, the major purpose of Experiment 2 was to assess retention of the detection skill over longer intervals.

In Experiment 1 we found only modest reductions in the frame size effects as a function of practice. However, subjects in Experiment 1 were given at most only four hours of practice. Therefore, a second purpose of Experiment 2 was to determine whether more intensive practice will lead to more dramatic changes in the frame size effects and, thus, to a greater degree of automaticity.

Two subjects were employed for Experiment 2. Each subject was given 12 one-hour sessions of detection training followed by a retention test 6 months after the training ended. One subject also received a second retention test 9 months after the first retention test (15 months after training ended). To verify further our findings from Experiment 1 with the prose passages, the subjects were also given pretest, posttest, and retention tests with the passage letter-detection task.

Method

Subjects

Two subjects were tested in this experiment. One subject (A.G.) was an undergraduate research assistant majoring in Psychology at the University of Colorado. She had had extensive experience testing subjects in experiments on cognitive psychology, including concurrent participation as an experimenter in Experiment 1. Although generally familiar with the stimulus configurations and tasks due to her role as experimenter, she remained in a separate room from the subjects during most of the training and testing sessions and did not view the stimulus displays during the presentation of the stimuli to the subjects. The

second subject (D.S.) had received his bachelor's degree from the University of Colorado within the previous year before training began. This subject had not been a psychology major and was not familiar with experimental psychology.

Design, Apparatus, and Procedure

All aspects of the apparatus and procedure are comparable to those for the analogous tasks of Experiment 1. In particular, the display duration was 1500 ms in both the detection training and passage letter-detection segments of the experiment. As in Experiment 1, the subjects completed a comprehension test for each passage.

Testing was conducted in three phases. Phase 1, the acquisition phase, consisted of a pretest on letter detection in prose followed by 12 sessions of intensive training carried out by A.G. within a 28-day period and by D.S. within a 40-day period. Phases 2 and 3 consisted of retention tests. Phase 2 consisted of one day of testing six months after the last day of training of Phase 1. Phase 3, which applied only to A.G., also consisted of a single day of testing nine months after the Phase 2 test day.

Phase 1: Acquisition. The first day of Phase 1 included only a pretest with Version 1 of the first letter-detection passage used in Experiment 1. Each of the remaining 12 days of Phase 1 included seven blocks of detection training comparable to that employed in Experiment 1 (i.e., uppercase letters, with uppercase H the target). The final day of Phase 1 also included two posttest passage letter-detection tasks, following the usual seven blocks of detection training. The first posttest letter-detection passage was Version 1 of the second passage used in Experiment 1. The second posttest letter-detection passage was one version of the t-detection passage used in previous studies (see, e.g., Proctor & Healy, 1985). Again, this passage was converted to all uppercase letters, and the target was uppercase T. Unfortunately, the data from A.G. for this passage were lost due to a computer malfunction, so the data from

this passage will not be reported for either subject.

Phases 2 and 3: Retention. Phases 2 and 3 each included a retention test for letter detection in prose, followed by seven blocks of the detection training task like that conducted in Phase 1. All stimuli were typed in uppercase, and uppercase H was the target in each task. Version 2 of the third passage used in Experiment 1 was presented in the prose letter-detection task of Phase 2. Version 2 of the pretest passage of Phase 1 was presented in Phase 3.

Results

Detection Training

Scoring procedures. The same scoring procedures were used as in Experiment 1. However, because only two subjects were tested in the present experiment, the factor of blocks (within days) rather than subjects, was treated as the random effect in two separate analyses of variance, one for each subject. Two types of analyses were conducted. The first type of analysis included only data from the first 12 sessions of training (Phase 1-Acquisition), with session and frame size as within-blocks factors. The second type of analysis included only data from the last session (Session 12) of acquisition training and the two retention sessions (Phases 2 and 3) for A.G. but only the one retention session for D.S., again with session and frame size as within-blocks factors. Thus, in the first type of analysis there were 12 levels for the session factor, whereas in the second type of analysis there were only 3 levels for A.G. and 2 levels for D.S. In both types of analyses there were three levels for the frame size factor (2, 4, and 16).

Accuracy data. Figure 2 shows the proportion of hits as a function of session and frame size for A.G. in the top panel and D.S. in the bottom panel. The standard error of the mean proportion of hits in Figure 2 is .008 for A.G. and .005 for D.S., as determined by analyses of variance. False alarms were also computed but were not analyzed further due to the low frequency of

occurrence during acquisition (mean = .04 for A.G. and .02 for D.S.).

For both subjects, the hit rates for the three frame sizes are quite different initially but converge as the training progresses, so that by the final (12th) session, the hit rates are at the ceiling for all three frame sizes and stay at the ceiling during the retention tests.

Insert Figure 2 about here

Response latency data. Figure 3 presents the means of the median response latencies for the hits as a function of session and frame size, again for A.G. in the top panel and D.S. in the bottom panel. The standard error of the mean response latencies in Figure 3 is 20 ms for A.G. and 22 ms for D.S., as determined by analyses of variance. In the analysis of variance for the acquisition period, there were significant main effects of session, $F(11, 66) = 37.93$, $p < .001$, for A.G., and $F(11, 66) = 11.20$, $p < .001$, for D.S., and frame size, $F(2, 12) = 204.92$, $p < .001$, for A.G., and $F(2, 12) = 922.84$, $p < .001$, for D.S., as well as a significant interaction between session and frame size, $F(22, 132) = 3.16$, $p < .001$, for A.G., and $F(22, 132) = 2.16$, $p < .01$, for D.S. In addition, trend analyses indicated that there was a significant linear Day X linear Frame Size interaction for A.G., $F(1, 6) = 26.17$, $p < .01$., but only a marginally significant interaction for D.S., $F(1, 6) = 3.79$, $p < .10$. As for the hit rate, the frame size effect diminished as the sessions progressed, but in this case the effect was not eliminated entirely at the end of acquisition training, presumably because the latencies had not reached their lowest possible level. Note that all three frame sizes converged for A.G., but only the smaller two frame sizes converged for D.S. In any event, the significant interaction does support the hypothesis that a degree of automaticity was achieved, but the difference in frame sizes at the last session, especially for D.S., suggests

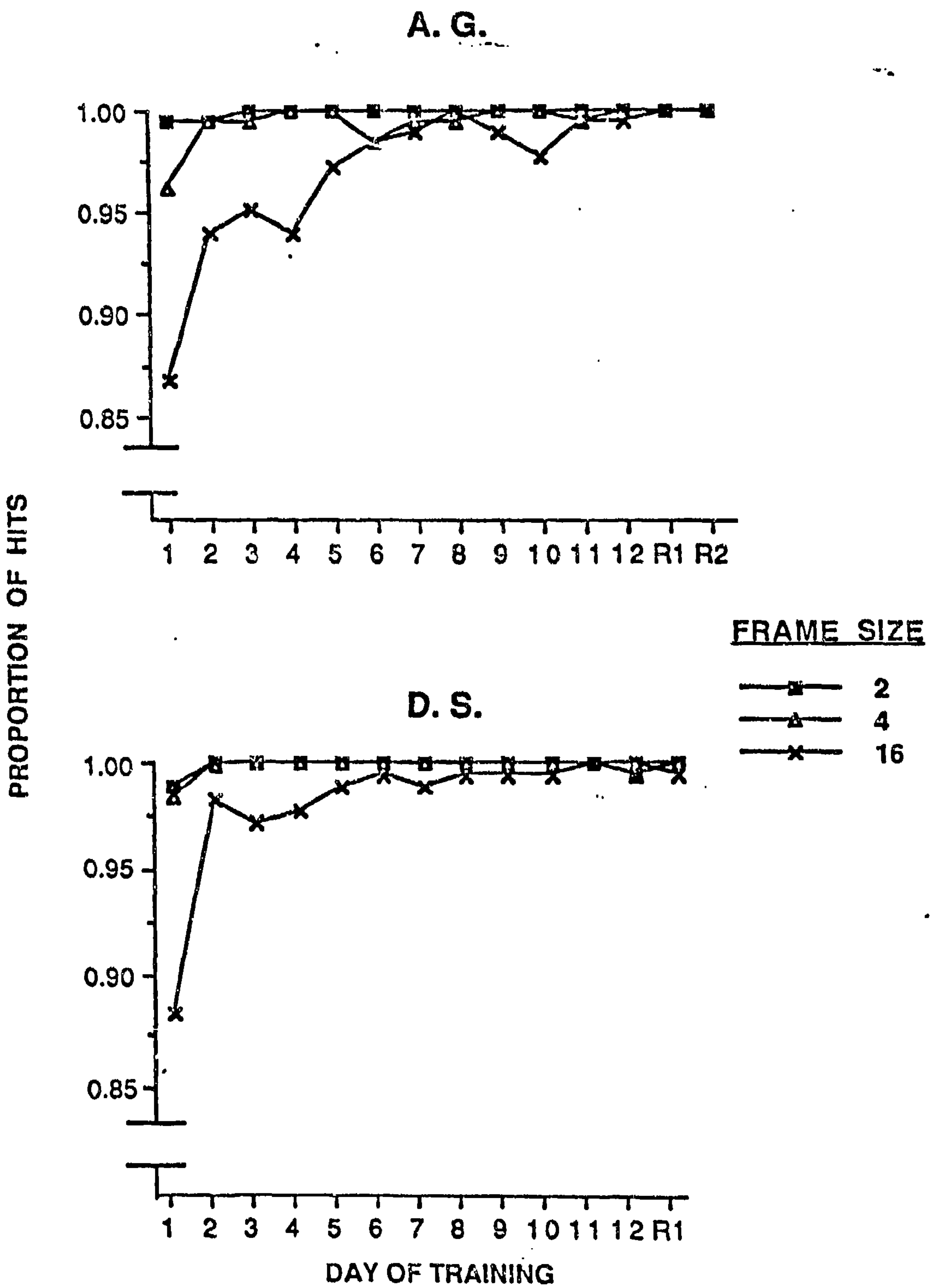


Figure 2. Mean proportion of hits as a function of day of training and frame size in Experiment 2.

that automaticity was not complete.

Insert Figure 3 about here

For the second analysis, which compared the last session of the acquisition period to the retention session(s), there was a significant main effect of frame size, $F(2, 12) = 43.82$, $p < .001$, for A.G., and $F(2, 12) = 96.15$, $p < .001$, for D.S., with the smallest latencies for Frame Size 2 and the largest for Frame Size 16. In addition, for D.S. there was no effect of session, $F(1, 6) < 1$, but for A.G. there was a significant main effect of session, $F(2, 12) = 8.68$, $p < .01$, with shorter latencies for the final acquisition session and the final retention session than for the initial retention session. Planned comparisons for A.G. revealed no significant difference between the first (final acquisition) and third (final retention) tests, but the average of the first and third tests did differ from that of the second (initial retention) test, $F(1, 6) = 13.70$, $p = .01$. Thus, for A.G. the final retention test 15 months after acquisition yielded performance comparable to that at the end of training, suggesting essentially no forgetting, although there was a significant performance decrement after the first 6-month interval for that subject. Alternatively, after the first 6-month delay, there was significant forgetting evident at the retention test, for A.G., but that test provided a reminder which boosted performance back to the level attained at the final acquisition session. In contrast, no forgetting was evident at the 6-month retention test for D.S.

<u>Letter Detection in Prose</u>

The scoring procedure for this task was the same as that used in Experiment 1. However, because this experiment included only two subjects, no statistical analysis was conducted.

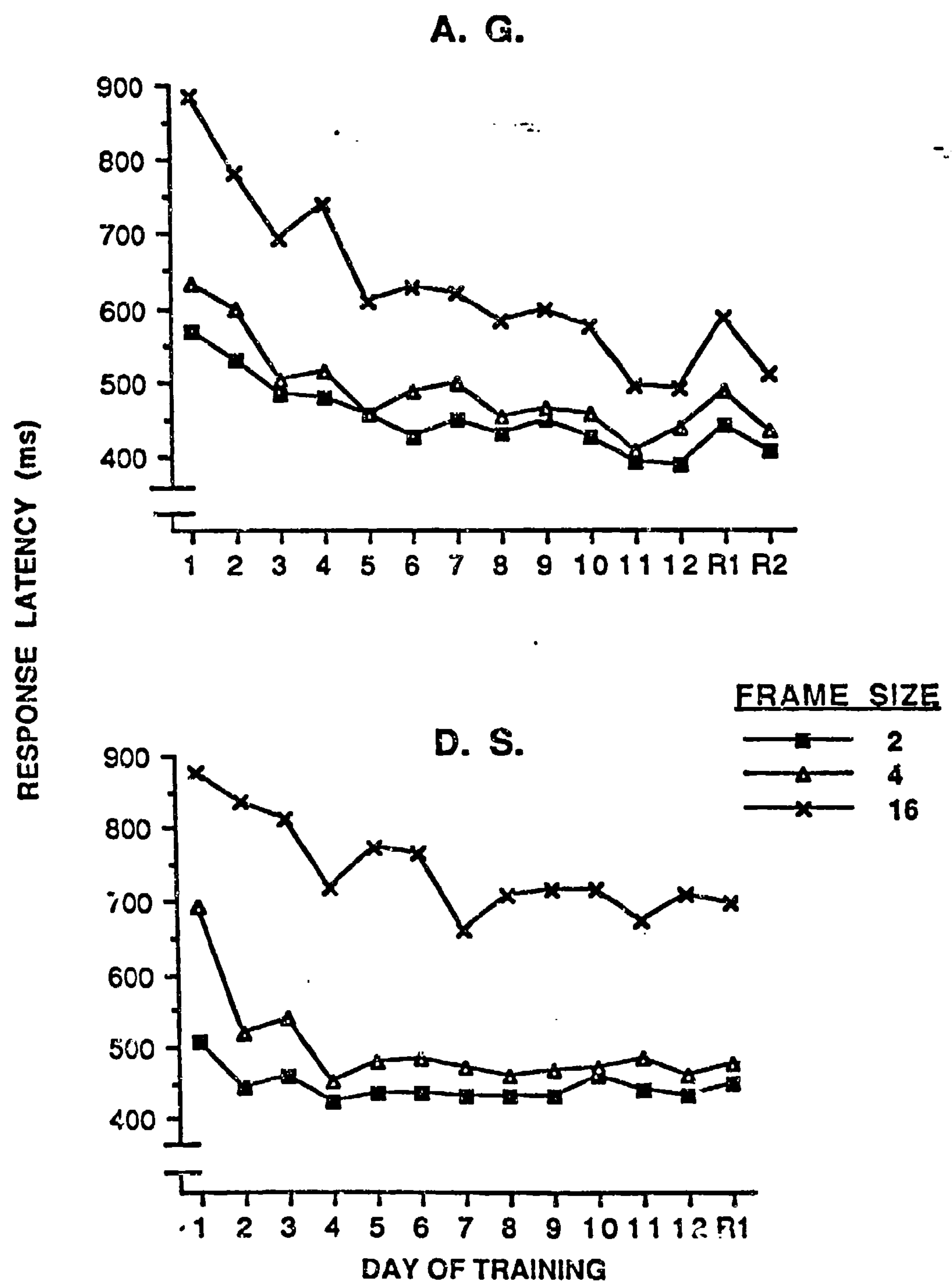


Figure 3. Average median response latency (in milliseconds) as a function of day of training and frame size in Experiment 2.

The proportion of hits and false alarms were computed for each of the tests of letter detection in prose. The false alarm rate was small for each of the tests (mean = .03 for A.G. and mean = .02 for D.S.).

The proportion of hits as a function of test, word type, and spelling are shown in Table 7. The most striking aspect of these data is the increase in hit rate from the pretest to the posttest. This improved performance is maintained during the retention tests for A.G. but not fully maintained for D.S. It is also noteworthy that the data from the pretest are consistent with both a word inferiority effect for THE (but not for other words) and a word frequency disadvantage. The data are generally inconsistent with a word frequency disadvantage and word inferiority effect for the posttest for both subjects and for the retention tests for A.G.

Insert Table 7 about here

Thus, the results for these two subjects are generally consistent with those from the previous experiments in two respects. First, the effects of training are large and are retained throughout long periods of disuse (in this case up to 15 months). Second, the standard word frequency disadvantage is eliminated at the posttest for both subjects and remains absent at the subsequent retention tests for A.G. (but not D.S.).

Discussion

The two subjects employed in this experiment elucidated two important effects of training suggested, but not clearly demonstrated, in Experiment 1, which involved a greater number of subjects but substantially less training and a shorter retention interval. First, performance, both in terms of speed and accuracy, improved dramatically as training progressed. Although it appeared from Experiment 1 that performance might have reached an asymptotic level with

Table 7

Proportion of Hits as a Function of Test, Word Type and Spelling: Experiment 2

Test	Word	A.G. Cor	A.G. Mis	D.S. Cor	D.S. Mis
Pretest					
	the	.72	.78	.61	.83
	other	.78	.72	.78	.83
Posttest H					
	the	.96	.92	.92	1.00
	other	.96	.98	.96	.93
Retention 1					
	the	.94	1.00	.80	1.00
	other	.86	.94	.94	.94
Retention 2					
	the	1.00	1.00	—	—
	other	1.00	1.00	—	—

Note. Cor = Correct; Mis = Misspelled.

as little as four days of training, it is clear instead that performance steadily increased throughout the 12-session training period. Although a frame size effect persisted for response latencies at the end of training, especially for D.S., the effect was substantially reduced for latencies and was eliminated for errors. Hence, it is clear that A.G. and probably D.S. became more automatic as a result of the training. Second, and most interesting, the large improvements in performance were maintained with essentially no decline over a six-month retention interval for D.S. and over a fifteen-month interval for A.G., with only one refresher training session intervening between the training and A.G.'s final retention test. This latter finding suggests that the perceptual skill of letter detection more closely resembles motor learning rather than verbal learning in its retention characteristics. The extremely large degree of retention evident here is surprising and certainly worth further exploration.

General Discussion

We can best summarize our findings by dividing them into three subsets, those concerning letter detection in prose, the role of automaticity, and long-term retention.

Letter Detection in Prose

In preliminary research (Healy, Fendrich, & Proctor, 1987) we found that the word frequency disadvantage was eliminated and the word inferiority effect was reduced after detection training. We eplicated this result in the present study, but we also found the same change in the pattern of detection errors when subjects were given no detection training but instead merely performed a pretest with the prose task. Moreover, these results wr uninfluenced by the degree of detection training. Hence, the changes we observed cannot be attributed to enhanced letter-level processing alone but also to exposure to a level of processing higher than the letter. These findings are consistent with those of

Kolers and Magee (1978) showing only a moderate amount of transfer from naming inverted scrambled letters to reading inverted text. Although higher-level processing is implicated, we can rule out passage familiarity as a factor, because, unlike previous studies (e.g., Healy, Oliver, & McNamara, 1987), the changes in the pattern of letter-detection errors occurred even when subjects were tested with a new passage never seen previously (cf., Levy, 1983; Levy & Begin, 1984; Levy et al., 1986). Hence, we propose that exposure to a pretest with the prose letter detection task enables subjects to change the focus of their attention from the word or phrase levels to the letter level, and thus, to detect target letters which would have otherwise been overlooked because they are in common words.

These findings are consistent with the basic assumption, discussed in the Introduction, that failures to detect letters in common, correctly spelled words result from interactions of processing at the word and letter level. As outlined earlier, this assumption leads to the prediction that enhancing processing at the letter level will change the pattern of letter-detection errors. Our results imply that such a change takes place only when practice occurs in the context of real words so that subjects can learn to focus their attention on the letter level. More specifically, with respect to the unitization hypotheses (see, e.g., Healy, Oliver, & McNamara, 1987), our findings suggest that the pull of the text caused by the comprehension processes can be weakened by practicing letter detection in a prose context, so that subjects can learn to continue processing a given word at the letter level even when that word has already been identified.

<u>The Role of Automaticity</u>

Although subjects did not show evidence for fully automatic responding in Experiments 1 and 2, all subjects did show clear signs of improvement with training, and the subjects in Experiment 2 showed dramatic improvements

approaching automaticity, especially A.G.

Our finding that the effect of frame size persisted even after extensive practice is consistent with the findings from an experiment by Rabbitt et al. (1979). These investigators independently manipulated both target (or memory) set size and display (or frame) size in their Experiment 2. There were three different numbers of targets (two, four, or eight) and three different numbers of letters in the display (two, four, or nine). After extensive practice (60,000 trials across 25 days) on a consistent-mapping visual search task, they found that the effect of target set size was eliminated, but the effect of display size remained.

Most crucially, the durability of the detection skill does not seem to depend on the development of automatic processing. We found in Experiment 1 essentially perfect skill retention for subjects given both limited and extensive training, although there was no evidence of automaticity for the limited training group. Further, we found in the same experiment that both groups of subjects maintained the loss of the word frequency disadvantage over a one-month retention interval.

Long-Term Retention

Our most interesting results concern the long-term retention of the letter-detection skill. Subjects showed essentially no forgetting of the skill that they had acquired even after relatively long retention intervals. This finding was most dramatic for the subjects of Experiment 2, who showed substantial improvements in performance as a result of training. D.S. showed no loss in the performance level achieved after a 6-month retention interval, and A.G. showed no loss after a 15-month retention interval with only a single refresher training session (the 6-month retention test) intervening between initial acquisition and the final retention test. Not only was there little forgetting of the detection skill, but the large change in performance on the

prose task (the elimination of the word frequency disadvantage) was generally (not for D.S.) maintained over a lengthy retention interval, even when the change was caused by an experience of a relatively short duration (i.e., reading a single pretest passage).

The previous study in the literature most closely related to our own is the one by Rabbitt et al. (1979). This study employed a visual search task like our own and similarly examined the training and subsequent retention of the search skill. In Experiment 1 of this study, 60 subjects were exposed to three days of training, with 1,000 trials per day (a total of 3,000 trials, similar to the 3,744 trials given to subjects in the extensive training group of our Experiment 1). They were then retested after retention intervals of two, four, or six weeks. For some subjects the retention test involved the same target and distractors as used in training, whereas for other subjects a transfer test involving new distractors was employed. Subjects showed improvements in response latencies as training progressed and no increase in response latencies after the two- and four-week intervals. There was a significant increase after the six-week delay, although the latencies in that case were shorter than those at the start of practice. Hence, the results pointed to substantial degrees of skill retention up to four weeks, as we found in Experiment 1. Further, the results indicated significant forgetting after a six-week delay, as we found in our Experiment 2 for A.G., but not for D.S., after a six-month delay. The superior retention we found in our Experiment 2 may be due to the fact that the subjects in that study were exposed to considerably more extensive practice (13,104 trials). Also, our finding no loss in A.G.'s performance after a 15-month interval suggests that a limited amount of refresher training can maintain the skill even if there is initially some forgetting. Thus, although our findings are not inconsistent with those from the earlier study, the work by Rabbitt et al. seems to have underestimated the remarkable durability of the

perceptual skill.

The negligible amount of forgetting found in our study of perceptual learning contrasts with the substantial forgetting found in traditional studies of learning (e.g., consider the rapid forgetting of three letters over an 18-second retention interval found by Peterson & Peterson, 1959). There are at least three interrelated distinctions between the present task and the traditional tasks, and any one or any combination of these distinctions may be responsible for the different patterns of forgetting. First, the tasks differ along the dimension which we will refer to as skill versus knowledge (see Bourne, Ekstrand, & Dominowski, 1971); others have labeled this dimension operational versus declarative knowledge (knowing how versus knowing that; Ryle, 1949), or procedural versus declarative memory (Anderson, 1983). The subjects in our study learned a skill, whereas the subjects in the more traditional studies of verbal learning acquired knowledge. Second, the tasks differ in terms of the memory systems distinguished by Tulving (1985). Subjects in our task were engaging the procedural memory system, whereas subjects in the traditional tasks were making use of episodic memory. Third, our task was an implicit memory test, as opposed to the explicit memory tests used in the traditional tasks (see, e.g., Graf & Schacter, 1985). Thus, the letter-detection task we studied was a skill involving an implicit test of the procedural memory system. Other examples of long-term retention with little forgetting have involved pursuit-rotor motor skills (e.g., Jahnke & Duncan, 1956), reading inverted text (e.g., Kolers, 1976), and the word-fragment completion test (e.g., Sloman, Hayman, Ohta, Law, & Tulving, 1988; Tulving, Schacter, & Stark, 1982). The pursuit rotor task seems to fall unambiguously in the domain of skill; the reading of inverted text seems to be a clear example of procedural memory; and the priming of word-fragment completion has been used as an implicit measure of memory.

All three of these distinctions point to the involvement of procedural memory as the crucial factor leading to stable memory representations. In agreement with the theoretical position put forth by Kolers and Roediger (1984), we propose that memory representations cannot be divorced from the procedures which were used to acquire them, and that the durability of memory depends criticially on the extent to which the learning procedures are reinstated at test. Implicit memory tasks like ours which require the direct storage and retrieval of procedures should, according to this argument, be acquired and maintained with much greater facility than explicit memory tasks which involve procedural memory more indirectly, such as those which have been categorized as involving knowledge or episodic memory. For example, in the standard list learning experiment, the memory coding procedures used by subjects to store the list are not easily retrieved or reinstated at the time of test, unless the subjects employ specific mnemonic procedures, such as the method of loci, the keyword method, or the chunking method learned by the expert S.F. (Ericsson & Chase, 1982). In contrast, the procedures used by our subjects during acquisition are easily reinstated during the retention test because the subjects are performing the same task (i.e., letter detection) in both instances. This characterization of memory is consistent with theories of transfer-appropriate processing (e.g., Bransford, Franks, Morris, & Stein, 1979) and encoding specificity (Tulving & Thompson, 1973), both of which postulate that memory performance will be best when the procedures required at the retention test match those employed during learning.

This emphasis on procedural memory not only provides an explanation for the substantial degree of retention we found of the detection skill in our study, but also helps explain another puzzling observation we have made. We have found that the pattern of errors on the prose letter detection task is influenced greatly by a previous experience with detection in prose but not by experience

with detection in scrambled letters. The lack of an influence in the latter case could be explained by proposing that subjects use qualitatively different procedures to detect letters in the two contexts.

References

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Bourne, L. E., Jr., Ekstrand, B. R., & Dominowski, R. L. (1971). The psychology of thinking. Englewood Cliffs, NJ: Prentice Hall.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), Levels of processing in human memory. Hillsdale, NJ: Erlbaum.

Crowder, R. G. (1976). Principals of learning and memory. Hillsdale, NJ: Erlbaum.

Drewnowski, A., & Healy, A. F. (1977). Detection errors on the and and: Evidence for reading units larger than the word. Memory & Cognition, 5, 636-647.

Dumais, S. T. (1979). Perceptual learning and automatic detection: Processes and mechanisms. Unpublished doctoral dissertation, Indiana University.

Ericsson, K. A., & Chase, W. G. (1982). Exceptional memory. American Scientist, 70, 607-615.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. Perception & Psychophysics, 8, 215-221.

Gardner, G. T. (1973). Evidence for independent parallel channels in tachistoscopic perception. Cognitive Psychology, 4, 130-155.

Graf, P., & Schacter, D.L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 501-518.

Healy, A. F. (1976). Detection errors on the word the: Evidence for reading units larger than letters. Journal of Experimental Psychology: Human Perception and Performance, 2, 235-242.

Healy, A. F., Conboy, G. L., & Drewnowski, A. (1987). Characterizing the processing units of reading: Effects of intra- and interword spaces in a letter detection task. In B. Britton & S. Glynn (Eds.), Executive control processes in reading (pp. 279-296). Hillsdale, NJ: Erlbaum.

Healy, A. F., & Drewnowski, A. (1983). Investigating the boundaries of reading units: Letter detection in misspelled words. Journal of Experimental Psychology: Human Perception and Performance, 9, 413-426.

Healy, A. F., Fendrich, D. W., & Proctor, J. D. (November, 1987). The effects of training on letter detection. Paper presented at the 28th Annual Meeting of the Psychonomic Society, Seattle, Washington.

Healy, A. F., Oliver, W. L., & McNamara, T. P. (1987). Detecting letters in continuous text: Effects of display size. Journal of Experimental Psychology: Human Perception and Performance, 13, 279-290.

Jahnke, J. C., & Duncan, C. P. (1956). Reminiscence and forgetting in motor learning after extended rest intervals. Journal of Experimental Psychology, 52, 273-282.

Kolers, P. A. (1976). Reading a year later. Journal of Experimental Psychology: Human Learning and Memory, 2, 554-565.

Kolers, P. A., & Magee, L. E. (1978). Specificity of pattern-analyzing skills in reading. Canadian Journal of Psychology, 32, 43-51.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. Journal of Verbal Learning and Verbal Behavior, 23, 425-449.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6, 293-323.

Levy, B. A. (1983). Proofreading familiar text: Constraints on visual processing. Memory & Cognition, 11, 1-12.

Levy, B. A., & Begin, J. (1984). Proofreading familiar text: Allocating resources to perceptual and conceptual processes. Memory & Cognition, 12, 621-632.

Levy, B. A., Newell, S., Snyder, J., & Timmins, K. (1986). Processing changes across reading encounters. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 467-478.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 88, 375-407.

McGeoch, J. A. (1942). The psychology of human learning: An introduction. New York: Longmans, Green, and Company.

McGeoch, J. A., & Melton, A. W. (1929). The comparative retention values of maze habits and of nonsense syllables. Journal of Experimental Psychology, 12, 392-414.

Naylor, J. C., & Briggs, G. E. (1961). Long-term retention of learned skills: A review of the literature. (ASD-TR-61-390). Wright-Patterson AFB, OH: Advanced Systems Division.

Peterson, L.R., & Peterson, M.J. (1959). Short-term retention of individual verbal items. Journal of Experimental Psychology, 58, 193-198.

Proctor, J. D., & Healy, A. F. (1985). A secondary-task analysis of a word familiarity effect. Journal of Experimental Psychology: Human Perception and Performance, 3, 286-303.

Rabbitt, P., Cumming, G., & Vyas, S. (1979). Improvement, learning and retention of skill at visual search. Quarterly Journal of Experimental Psychology, 31, 441-459.

Ryle, G. (1949). The concept of mind. London: Hutchinson.

Schneider, W., & Fisk, A. D. (1983). Attention theory and mechanisms for skilled performance. In R. A. Magill (Ed.), Memory and control of action (pp. 119-143). New York: North-Holland Publishing Company.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. Psychological Review, 84, 1-66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Sloman, S.A., Hayman, C.A.G., Ohta, N., Law, J., & Tulving, E. (1988). Forgetting in primed fragment completion. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 223-239.

Treisman, A., & Paterson, R. (1984). Emergent features, attention, and object perception. Journal of Experimental Psychology: Human Perception and Performance, 10, 12-31.

Tulving, E. (1985). How many memory systems are there? American Psychologist, 40, 385-398.

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 336-342.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80, 352-373.

Author Notes

Portions of this article were read at the 28th Annual Meeting of the Psychonomic Society, Seattle, Washington, November 6, 1987.

This research was supported in part by National Science Foundation Grant Number BNS-8025020, United States Air Force Human Resources Laboratory Contract Number TL5744-022-001, and United States Army Research Institute Contract Number MDA903-86-K-0155 to the Institute of Cognitive Science at the University of Colorado, by Biomedical Research Support Grant RR07013-16 awarded to the University of Colorado by the Biomedical Research Support Grant Program, Division of Research Resources, National Institutes of Health, by a Council on Research and Creative Work Grant-in-Aid from the University of Colorado, and by National Institute of Mental Health Training Grant Number MH14617-08 to the University of Colorado. Alice F. Healy was supported by a Faculty Fellowship from the University of Colorado and a James McKeen Cattell Fund Sabbatical award during the preparation of this article.

We are indebted to Richard M. Shiffrin for help designing a preliminary experiment and for many stimulating discussions about this research, to Ernest Mross for help with computer programming, to Antoinette Gesi for help with testing subjects and data analysis, to Lyle E. Bourne, Jr. for many valuable suggestions concerning these experiments, and to Henry L. Roediger III, Keith Rayner, Tom Carr, and three anonymous reviewers for helpful comments about this report.

Correspondence concerning this article should be addressed to Alice F. Healy, Department [illegible] Psychology, Campus Box 345, University of Colorado, Boulder, Colorado [illegible]09.

APPENDIX C

THE DISAPPEARANCE OF A WORD INFERIORITY EFFECT: STRATEGY SHIFT OR PERCEPTUAL EFFECT?

In the common pattern matching or letter detection tasks involving single patterns or words, stimulus familiarity typically facilitates performance. However, when the stimuli occur in a continuous prose format, familiarity can have the opposite effect. Healy and her associates have repeatedly shown that when subjects are required to search for a target letter such as H, while reading for comprehension, targets located in high frequency words such as THE, are detected less frequently than targets in other, lower frequency words such as THY, or in misspelled words such as THD.

Drewnowski & Healy have proposed a set of unitization hypotheses to explain these effects, according to which two parallel processes, word processing and letter processing, proceed until a word is identified. Once word identification occurs, all processing ceases. In the case of very familiar, high frequency words such as THE, word identification may be completed before all letters have been identified. As a result, detection of target letters in these words is impaired, producing a word frequency disadvantage.

In a recent series of studies reported this fall at the meeting of the Psychonomics Society, Alice Healy, David Fendrich and I found that subjects no longer performed worse with high frequency words compared to lower frequency words on a second letter detection task following automaticity training with the target letter. However, neither did subjects given automaticity training with a neutral target, a number. For both groups, the hit rate for the target letter H in the word THE dramatically improved relative to the hit rate for the letter H in other words.

(Presented at the 34th Annual Meeting of the Southeastern Psychological Association, March 31-April 2, 1988, New Orleans, LA.)

Why should the effect disappear or even reverse when performing the task a second time? Subsequent experiments ruled out a number of procedural explanations, and demonstrated that the automaticity task is irrelevant to the reduction of the word frequency disadvantage; performing the letter detection task a second time is all that is necessary. Therefore, we are left with two likely possibilities. First, the effect might disappear simply because subjects come to realize that the target always is present in the word THE, and at that point, change their strategy to one of responding whenever the word THE is seen. A comparison of letter detection in the first half of the initial passage in the automaticity study did reveal a stronger word frequency disadvantage than that present in the second half of the passage. This is at least consistent with a strategy shift explanation, but it does not rule out an alternative.

This second possibility is that even limited experience with the letter detection task is sufficient to shift the emphasis from attention to the word or phrase level, to attention to the letter level. Consistent with this explanation is the finding reported by Proctor & Healy (1985) that the disadvantage of word frequency is reduced when instructions emphasize letter detection rather than text comprehension. Again, however, these data are not conclusive.

The studies I will present today were conducted to investigate the disappearance of the word frequency disadvantage with specific emphasis on the first of the two proposed explanations; that is, that the effect disappears because subjects begin to look specifically for the word THE.

In Experiment 1, 48 subjects read a prose passage printed on paper and marked instances of the letter H. The 483-word passage included 72 test words containing H, 36 of which were the word THE. Two instructions conditions were

compared, one (the "standard" condition) in which subjects were given standard instructions, and another (the "hint" condition) in which subjects were given an explicit hint. All subjects were instructed that their primary task was to read for comprehension in preparation for subsequent questioning, and that as a secondary task they should mark each instance of the target letter H seen. Subjects in the HINT group received enriched instructions that included the following statement: "A hint for you is to notice that the letter H always appears in the word THE which is a very frequent word. Thus, you should try not to forget the word THE."

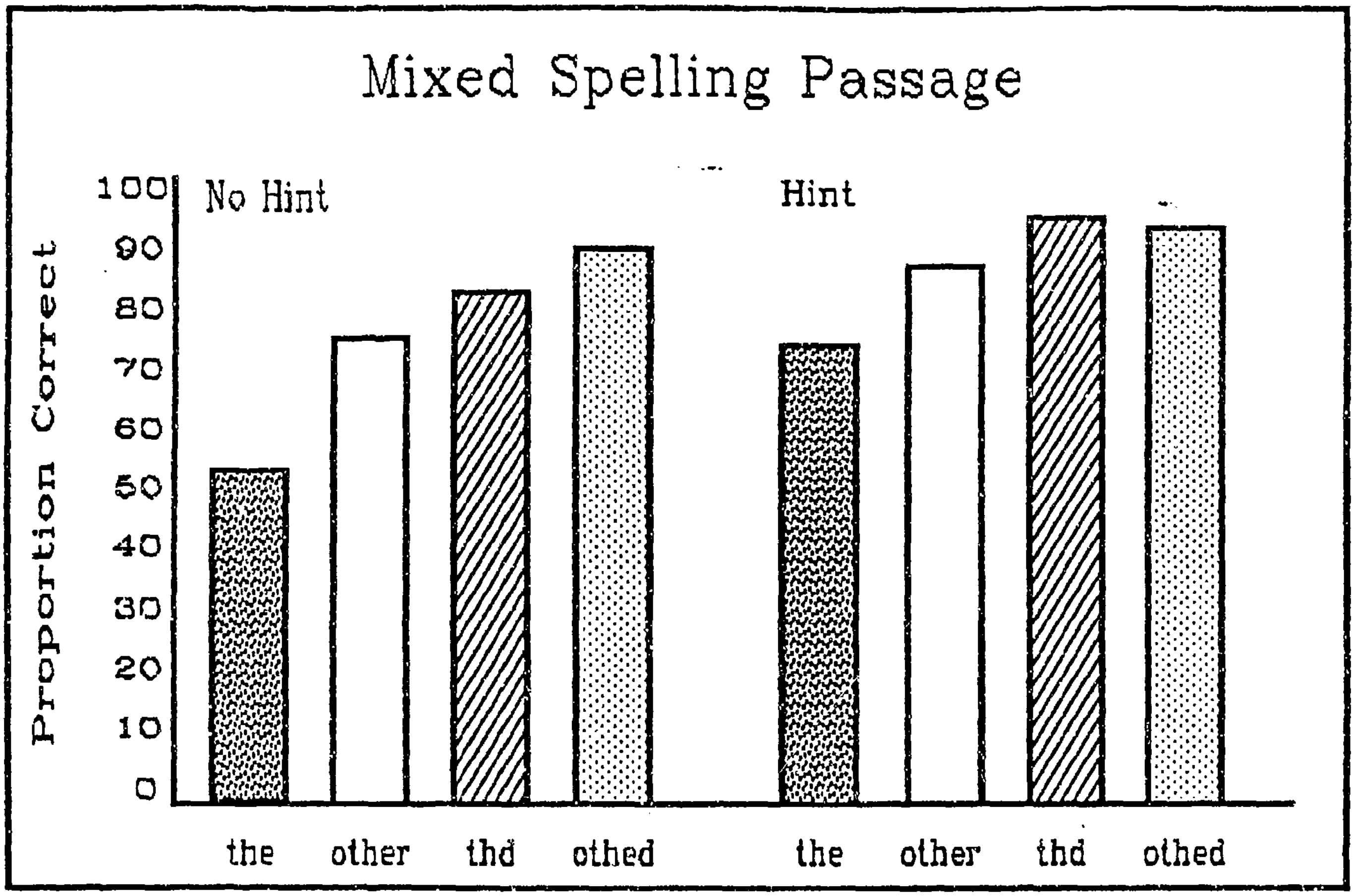
Additionally, two types of passages were employed, one in which all words were correctly spelled, and one in which half of each type of the target words and 12 filler words were misspelled. The misspelling manipulation has been a standard component of previous research involving the unitization hypothesis. Also, the frequent appearance of misspelled THE (always spelled THD) might serve to alert the subject to the fact that THE contains the target letter. Therefore, if the misspelling of the word THE is crucial, then the word frequency disadvantage should be reduced for the misspelled passage relative to the correctly-spelled passage, even without the hint instructions.

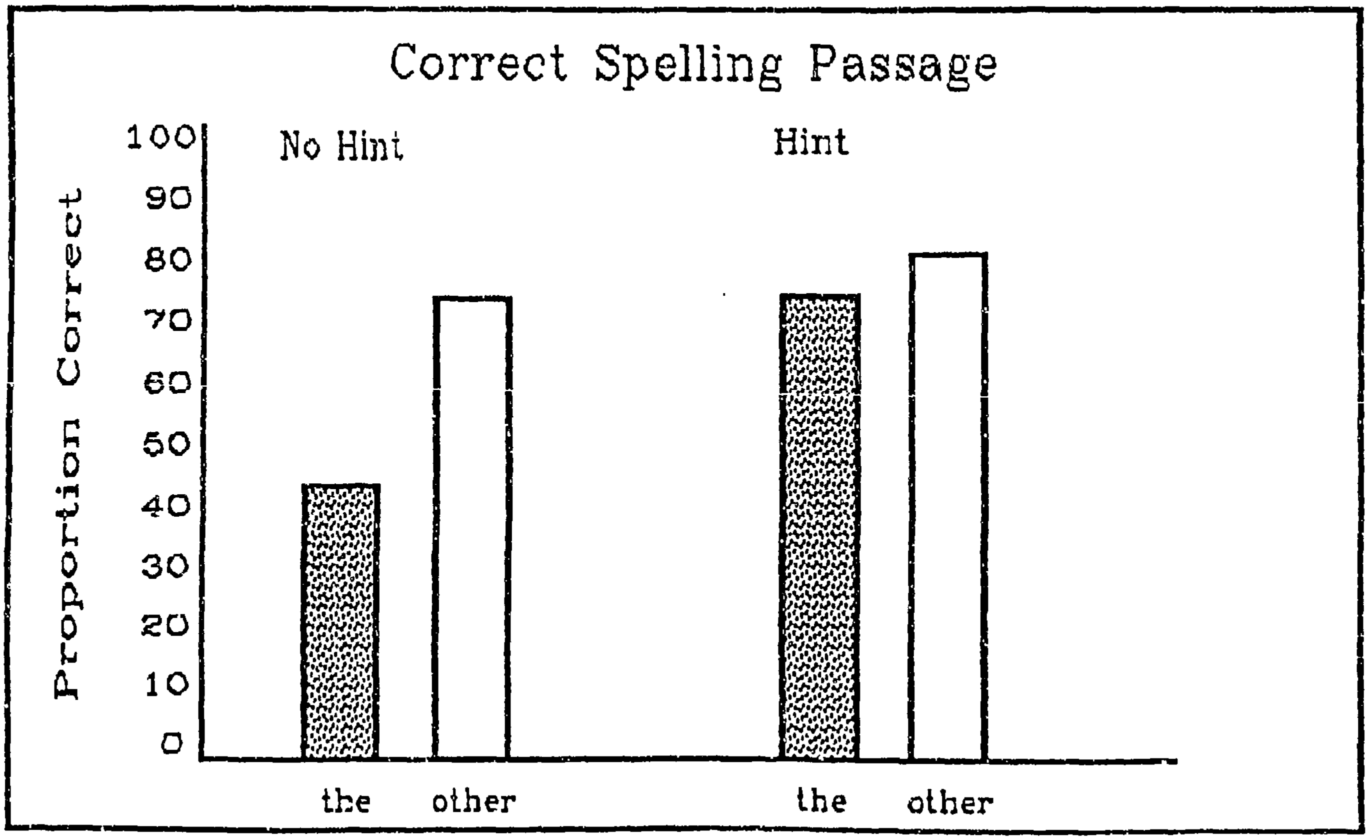
Following completion of the letter detection task, subjects were given 8 multiple-choice questions about the content of the passage. The overall proportion of correct responses was quite high, .76, and no significant effects were obtained. Therefore, the hint instructions and misspelling of words did not influence comprehension performance.

The proportion of hits in the letter detection task were computed as a function of test word type (THE vs OTHER) and spelling. Group means are shown in FIGURE 1.

PRESENT FIGURE 1 HERE

Figure 1


Mixed Spelling Passage
Proportion Correct
100
90
80
70
60
50
40
30
20
10
0
No Hint
Hint
the
other
thd
othed
the
other
thd
othed


Correct Spelling Passage
Proportion Correct
100
90
80
70
60
50
40
30
20
10
0
No Hint
Hint
the
other
the
other

As can be seen in FIGURE 1, and confirmed by the analyses of variance, the word frequency disadvantage that is clearly apparent under the standard instructions, was reduced under the hint instructions. Looking first at the correctly spelled words in both passage formats, we found that subjects made more hits on other words than on the word THE (a word frequency disadvantage), and that subjects in the hint group made a greater proportion of hits in general than those given the standard instructions. More importantly, with the HINT instructions, hits on THE increased more than did hits on other words. Note, however, that the hint did improve performance on the other words to some extent, and that although the hint instructions did reduce the magnitude of the word frequency disadvantage, it did not reverse or eliminate that effect.

As expected based on the first analysis, when both correctly and incorrectly spelled words from the mixed spelling passage were examined, we again found the word frequency disadvantage, and with the HINT instructions, a superior level of performance overall. Also, as is commonly found in this type of letter detection task, a greater proportion of hits were made on misspelled words than on correctly spelled words, and the effect of spelling was greater for THE than for other words. The presence of misspelled words did not alter the pattern of responsees in general, but the near ceiling levels of performance on misspelled words prevented an effect of the hint on the misspelled words.

The results of Experiment 1 indicate that subjects given a very specific hint to look for the word THE do, indeed, have a highter hit rate for targets in the correctly-spelled word THE, and thus, a smaller word frequency disadvantage, regardless of passage format. This is, of course, consistent with the strategy shift hypothesis. However, Experiment 1 does not prove that

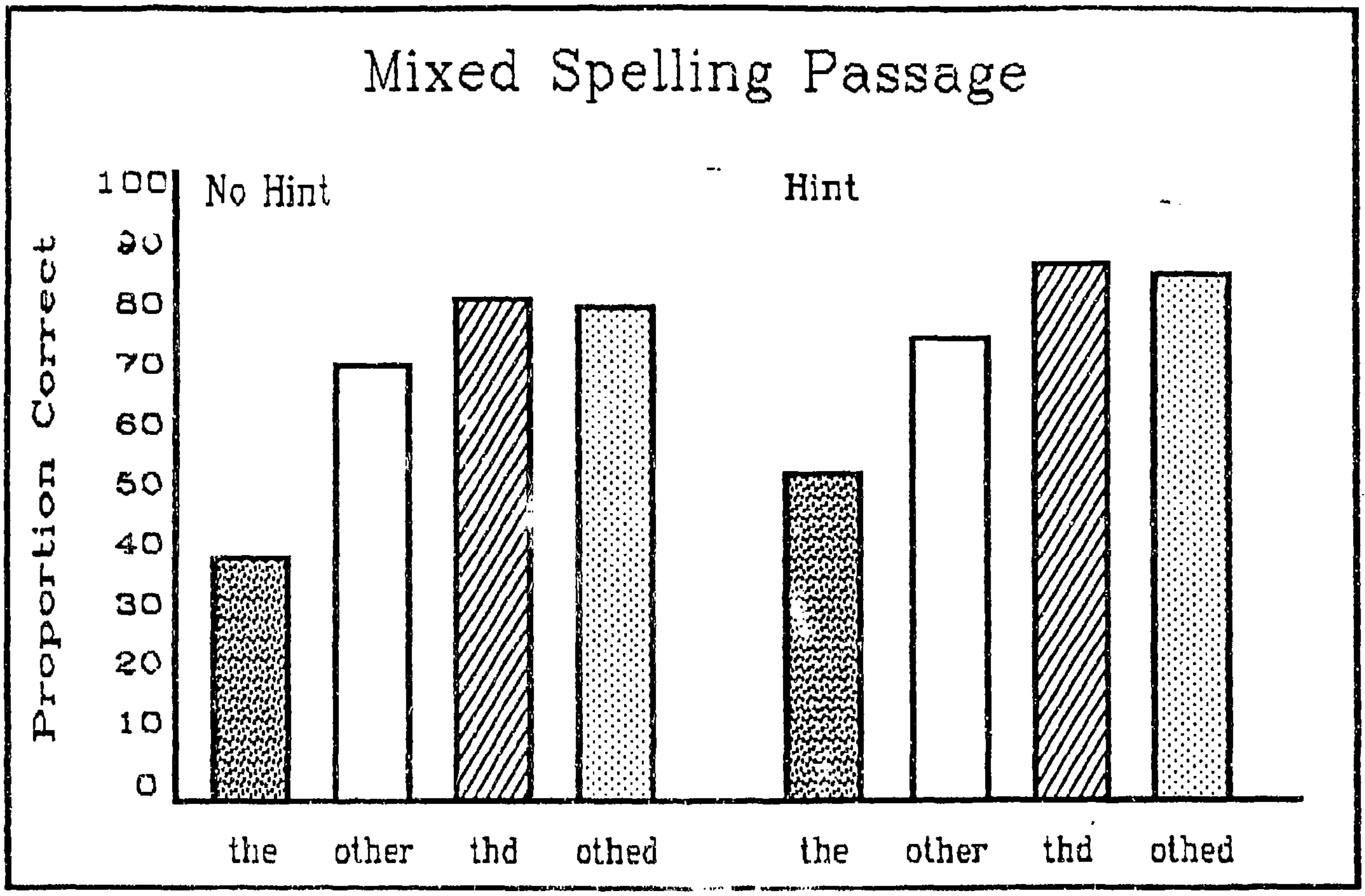
subjects are actually adopting this strategy. Also, in previous studies the word frequency disadvantage was entirely eliminated, or even reversed, whereas in Experiment 1 the effect was merely reduced. Therefore, a strategy shift might not be the only factor responsible for the loss of the word frequency disadvantage in previous studies.

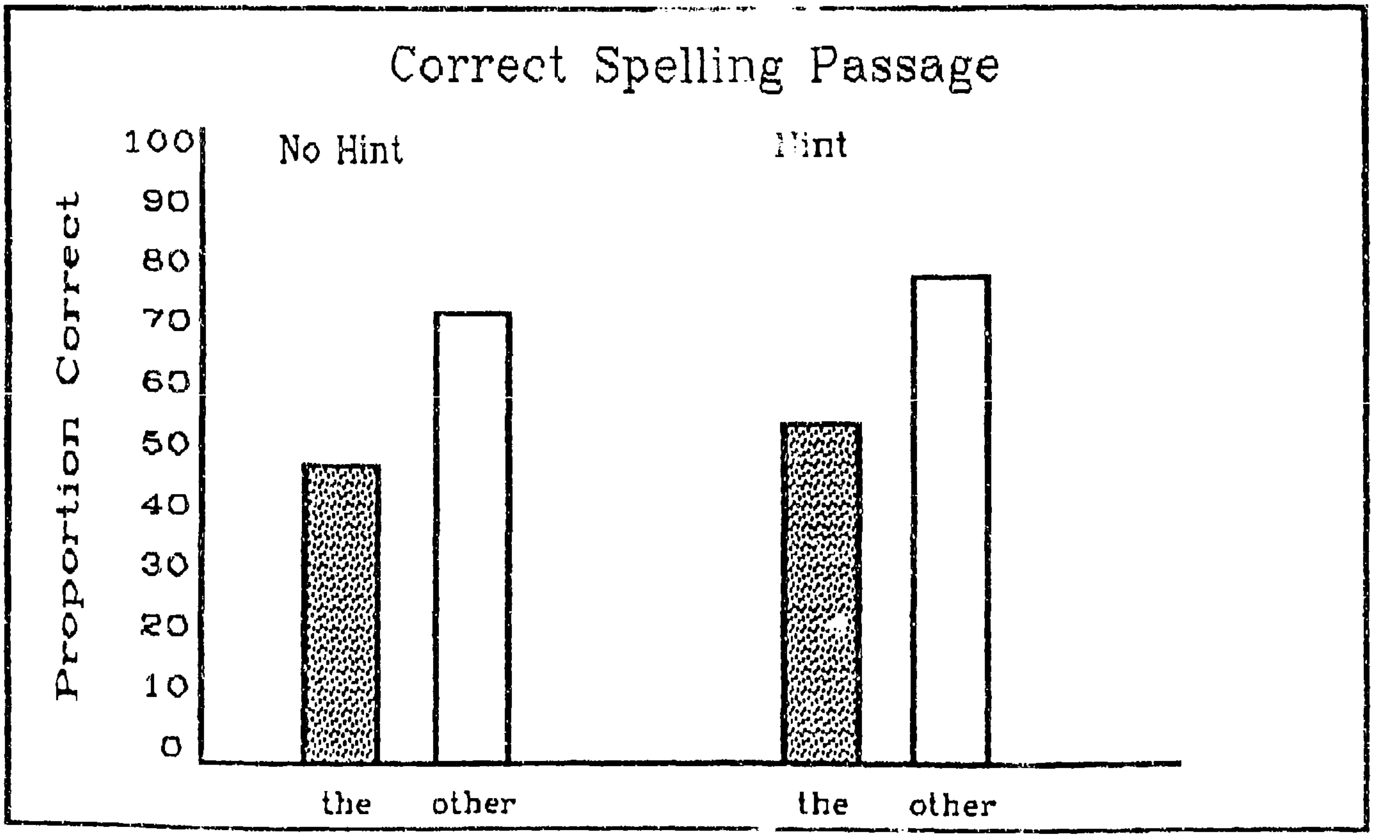
For example, the explicit instructions of Experiment 1 might have changed how subjects performed the detection task in ways other than just adopting a "look for THE" strategy. Stating that the letter H is in the word THE and emphasizing the need to try not to miss the word THE could also have shifted more attention to the letter level and encouraged greater attention to the letter detection task in general. This is consistent with the general improvement in the hit rate found in Experiment 1. Therefore, in Experiment 2 we modified the procedure somewhat to reduce the liklihood of changing the general nature of the detection task and to obtain more specific information concerning the strategies used by subjects.

The tasks and stimuli for Experiment 2 were identical to those of Experiment 1 except in three respects. First, the hint instructions were made less explicit. Subjects were not directly told that the word THE always contains the target letter. Instead, for the hint group the example text segment given with the instructions included the word THE, whereas the standard example did not. Second, reading time for the passage was measured for each subject. And third, following completion of the letter detection task, subjects were given a questionaire asking for information about their strategy and other aspects of the task or their performance that they noticed. The mean proportion hits for Expeirment 2 are shown in FIGURE 2.

PRESENT FIGURE 2 HERE

Figure 2


Mixed Spelling Passage
Proportion Correct
100
90
80
70
60
50
40
30
20
10
0
No Hint
Hint
the
other
thd
othed
the
other
thd
othed


Correct Spelling Passage
Proportion Correct
100
90
80
70
60
50
40
30
20
10
0
No Hint
Hint
the
other
the
other

An analysis of variance for the correctly spelled words in both passage formats indicated only a significant main effect of word type. As before, targets were detected less often in the word THE than in other words. Note that the subtle hint and the passage format had no effect.

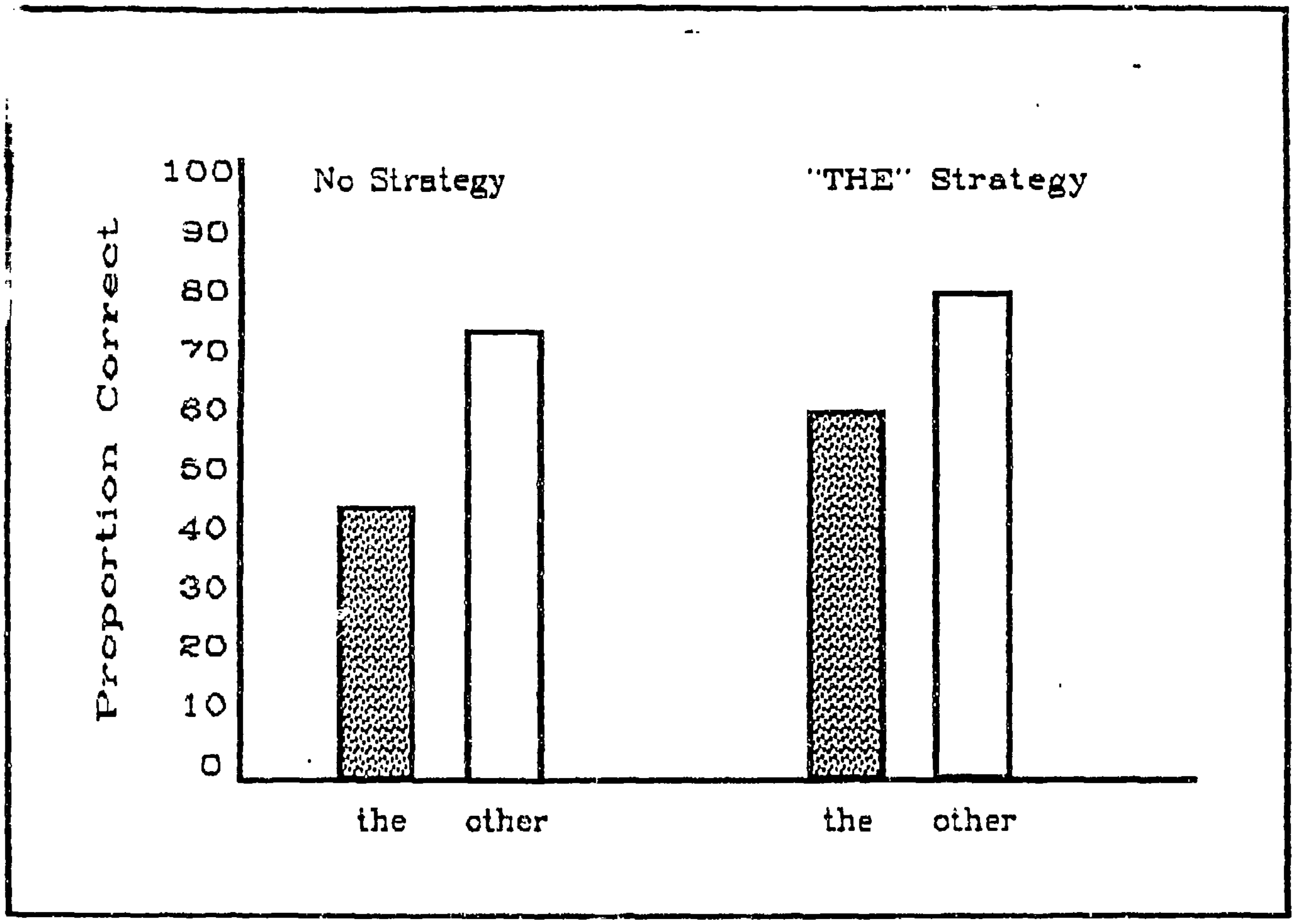
With the mixed spelling passage data, the standard pattern was again obtained. Significant main effects of word type and spelling and the Word Type X Spelling interaction were present. As in the analysis of the correctly spelled target words, the subtle hint had NO effect.

The subtle hint also had no effect on comprehension scores, nor did passage format. When reading times were analyzed, however, the subtle hint did lengthen the time spent reading by approximately 45 seconds or 15%. Somewhat surprisingly, whether or not the passage included misspellings did not affect reading time significantly.

Because no effect of the subtle hint was obtained, the results of the questionaire can offer less insight into the possible influence of subject strategies than had been anticipated. The results do, however, indicate that when the word frequency disadvantage is obtained, few subjects report using the strategy of looking for THE. When subjects were asked the question, "Which strategies or procedures, if any, did you use to detect the letter H?" only 8 out of 48 subjects reported looking for THE, and this strategy was reported more often by subjects in the standard instructions group than by those in the hint group, specifically, 5 subjects versus 3.

A final analysis was performed in which the subjects' stated strategy was used to form two comparison groups. Here, passage format and instructions conditions were not considered as a separate factors, and only data for correctly spelled words were analyzed. The means for the regrouped data are shown in FIGURE 3.

Figure 3



No Strategy
"THE" Strategy
Proportion Correct
100
90
80
70
60
50
40
30
20
10
0
the
other
the
other

SHOW FIGURE 3 HERE

Although the means in Figure 3 suggest that subjects who use the strategy of looking for THE have higher hit rates in general than other subjects and that this strategy reduced the magnitude of the word frequency disadvantage from a difference of .27 to .18, the analysis indicated no significant effects of strategy. Only the main effect of word type was significant. This does not support the hypothesis that the word frequency disadvantage is reduced by a strategy of looking for THE, and it leaves open the possibility that the reduction of the effect in Experiment 1 was due to other, more general shifts in strategy toward the letter detection task. Of course, without additional subjects in the THE strategy group, these data are far from conclusive.

Based on these data and those of Experiment 1, at this time we must conclude that a strategy shift to looking for the word THE <u>does</u> remain a likely factor in the reduction of the word frequency disadvantage. Explicit hints suggesting this strategy <u>do</u> significantly reduce the effect, and subjects receiving standard instructions sometimes adopt a "look for THE" strategy spontaneously. However, subjects who did adopt this strategy in Experiment 2 did not <u>clearly</u> show a reduced word frequency disadvantage, and the effect was never completely eliminated as has occurred in previous studies. Therefore, at this time, we cannot rule out other, perhaps perceptual, influences. We are currently conducting research designed to address this possibility.

APPENDIX D

COGNITIVE OPERATIONS AND THE GENERATION EFFECT

Robert J. Crutcher and Alice F. Healy
University of Colorado

Abstract

The present experiments were designed to test a cognitive operations hypothesis of the generation effect, or memorial advantage for material that is generated rather than simply read. In Experiments 1 and 2 the internal or external locus of cognitive operations was manipulated independently of the internal or external locus of stimulus production. In Experiment 3 the internal or external locus of cognitive operations was manipulated independently at the time of studying the material and at the time of the retention test. Subjects performed simple multiplication problems with the answers supplied either by themselves or the experimenter and with the calculations performed either by themselves or by another agent. A highly significant retention advantage was found for the tasks requiring internal multiplication operations at study, but there was no main effect for the distinction between subject and experimenter supplied answers or for the difference between internal and external multiplication operations at test. A general explanation for these results in terms of cognitive operations is considered as well as a more specific explanation in terms of using such operations as retrieval cues.

A growing number of experiments have demonstrated a distinct retention advantage for material that is generated by an individual rather than simply read. In these experiments, the stimuli are often pairs of words presented to subjects under two conditions: read and generate. In the read condition, a pair of words is presented and subjects read the pair aloud. In the generate condition, a word pair is presented with the first word intact and the second missing one or more letters; subjects must then generate the second word of the pair using the first word as a context. In the original experiments by Slamecka and Graf (1978; see also Jacoby, 1978), subjects were provided with five different contexts or rules for generating the target words, including: rhyming (e.g., save-cave); associate (e.g., lamp-light); category (e.g., ruby-diamond); opposite (e.g., long-short); and synonym (e.g., sea-ocean). Regardless of the generate context or production rule and regardless of the specific retention task (e.g., free recall, cued recall, or recognition), subjects consistently showed better retention for the generated items versus the items that were simply read.

Since these first experiments, this generation effect, as it has been called, has been replicated with a wide range of materials and a variety of generate rules (although the effect has been reversed under some circumstances; see Jacoby, 1983). The effect has been demonstrated for words (Jacoby, 1978; Slamecka & Graf, 1978; Donaldson & Bass, 1980; McFarland, Frey, & Rhodes, 1980; Glisky & Rabinowitz, 1985; and Nairne, Pusen, & Widner 1985), sentences (Graf, 1980), meaningful bigrams (Gardiner & Hampton, 1985), and numbers (Gardiner & Rowley, 1984; and Gardiner & Hampton, 1985), using semantic, orthographic, rhyming, and other generate rules. The effect has even been demonstrated for words that subjects tried but failed to generate (Slamecka & Fevreiski, 1983).

The most notable limiting factor on the effect has been the meaningfulness of the items generated: Until recently, no generation effect had been demonstrated for nonwords (McElroy & Slamecka, 1982), anomalous sentences (Graf, 1980), meaningless bigrams (Gardiner & Hampton, 1985), or nonunitized numbers (Gardiner & Hampton, 1985)—a nonunitized number would be 2 8 instead of 28, so the subject must say "two eight" rather than "twenty-eight."

Despite the robustness of the generation effect, explanations for it have not enjoyed any great consensus. However, of the many explanations proposed, two classes have appeared repeatedly. The first class consists of explanations that appeal to semantic memory involvement (McElroy & Slamecka, 1982; Slamecka & Fevreiski, 1983; and Nairne et al., 1985). The second class consists of explanations that attribute the effect to the process of generation itself. For example, generation requires increased arousal (Jacoby, 1978) or increased cognitive effort (Griffith, 1976; McFarland et al., 1980).

Of the explanations that implicate semantic memory, the most popular has been the "lexical activation" hypothesis (McElroy & Slamecka, 1982; Nairne et al., 1985; Payne, Neely, & Burns, 1986) which specifies that the generation effect depends upon the lexical status of the items generated. In generating a word, a subject searches semantic memory and as a result of this search activates related semantic features that can later serve as retrieval cues to access the generated word. In simply reading an item these related semantic features are not activated and hence cannot aid retrieval of the target item. Thus, according to this explanation, the retention advantage afforded by generation is not really due to generation per se, but rather to the enhanced activation of related semantic features which generate tasks induce. The strongest support for this position has been the lack of a generation effect for nonwords (McElroy & Slamecka, 1982), because nonwords presumably lack related

semantic features. Further support for this hypothesis comes from an experiment in which it was demonstrated that even words which subjects had attempted but failed to generate were more likely to be recognized on a subsequent test than words which subjects had simply read (Slamecka & Fevreiski, 1983). Apparently, even an unsuccessful generation attempt activates enough related semantic features to aid in a recognition task whereas a read task does not.

Despite its popularity, recent experiments suggest the lexical activation hypothesis may be too narrow in limiting the effect only to items represented in the subjective lexicon. Gardiner and Hampton (1985), for example, have demonstrated a generation effect for nonlexical items (i.e., items that did not correspond to single words) such as unitized numbers (e.g., 28) and meaningful bigrams (e.g., ET) as well as for familiar word pairs (e.g., cheese cake generated using the rule: a cake made of cheese). However, they have found no generation effect for nonunitized numbers (e.g., 2 8), meaningless bigrams (e.g., EO), or unfamiliar yet meaningful word pairs (e.g., tomato cake). From these results, they have argued that the generation effect does not depend upon the lexical status of the generated item but upon the correspondence of the generated item with some familiar concept in memory. Like the lexical activation hypothesis, though, Gardiner and Hampton's explanation makes the generation effect dependent on existing semantic structures—that is, the generated items must be represented as a functional unit in memory.

Another class of explanations has attempted to attribute the generation effect not to some activated existing mental structure but to the process of generation itself. For example, generation requires greater involvement of the self schema or self system (e.g., Banaji & Greenwald, 1984; Greenwald, 1981); or generation induces greater arousal and heightened arousal leads to increased retention (Jacoby, 1978). Another explanation suggests that generation

increases retention because it requires greater cognitive processing or mental effort (e.g., McFarland et al., 1980). Previous studies have demonstrated a relationship between cognitive effort and retention (Auble & Franks, 1978; Tyler, Hertel, McCallum, & Ellis, 1979; McFarland et al., 1980; McDaniel, 1981; Ellis, Thomas, & Rodriguez, 1984; but see Zacks, Hasher, Sanft, & Rose, 1983). Griffith (1976) attempted to measure the amount of processing effort in a generation task. He found that latencies on a secondary reaction time task were longer for a task in which subjects generated their own sentence from two experimenter-supplied words compared to a task in which subjects simply repeated or read a sentence in which the two words were already included, and retention for the pair of words was much better for the sentence generation condition. The longer latencies were an index of the greater cognitive capacity or attention required by the generate task. Because the argument against an effort explanation of the generation effect has relied on the lack of an effect for nonwords (McElroy & Slamecka, 1982), a recent finding that the generation effect can be obtained for nonwords (Nairne & Widner, 1987) suggests a reconsideration of a cognitive effort explanation.

Although the effort and lexical activation hypotheses have been seen as providing opposing accounts of the generation effect (see, e.g., McElroy & Slamecka, 1982), if viewed from a different perspective they can be seen as complementary. According to both the lexical activation hypothesis and the revised formulation proposed by Gardiner and Hampton (1985), the lexical status of a stimulus or its existence as a familiar concept in memory is important only because it allows for the mental activation of associated information. Likewise, it is the mental activity involved in generation which is seen as essential by the effort explanation. By both hypotheses, then, the crucial aspect of the generation effect is the inducement of auxiliary mental processes

or cognitive operations by the subject. If it is assumed that the generation effect is due to the activation in the subject of auxiliary cognitive operations, then a task leading the subject to perform such cognitive operations but not necessarily overt generation of an item may show equivalent retention to a generate task. Likewise, a task involving overt generation by the subject but no auxiliary cognitive operations may not result in any better retention than a read task. In other words, according to this formulation it is not essential that the subject rather than the experimenter generate or produce the stimulus, but rather it is essential that the subject rather than another agent engage in the auxiliary cognitive operations linking the stimulus to other information stored in memory. That is, the distinction between internal versus external stimulus production is not as important as the distinction between internal versus external cognitive operations.

In order to test this cognitive operations hypothesis, we have devised an experimental paradigm which allows for the orthogonal variation of locus of stimulus production (internal or external) and locus of auxiliary cognitive operations (internal or external). More specifically, we adapted the procedures used by Gardiner and Rowley (1984), in which subjects are given simple multiplication problems and required to remember their answers. In order to vary the locus of stimulus production, the answers are either given by the experimenter (external) or replaced by question marks so that they must be provided by the subjects (internal). Further, in order to vary the locus of auxiliary cognitive operations, the subjects are either required to perform the multiplication operations themselves (internal) or not so required (external). Thus, four tasks are included, which will be designated the "read," "generate," "verify," and "calculate" tasks. The read and generate tasks are equivalent to those employed by Gardiner and Rowley (1984). The read task involves both

external stimulus production and external multiplication operations, whereas the generate task involves both internal stimulus production and internal multiplication operations. The verify and calculate tasks are new conditions critical for testing our hypothesis. The verify task involves external stimulus production but internal multiplication operations. Specifically in this task (as in a procedure used by Donaldson & Bass, 1980), subjects are given the problem with its answer but are required to verify that the answer is correct. In contrast, the calculate task involves internal stimulus production but external multiplication operations. In particular, subjects in this task must provide the answer to the problem but they are told to use a calculator rather than perform the arithmetic themselves. The cognitive operations hypothesis yields the prediction that retention on the verify and generate tasks would be superior to that on the read and calculate tasks, because the former two tasks involve internal multiplication operations whereas the latter two tasks involve external multiplication operations. In contrast, no difference is expected between the generate and verify tasks or between the calculate and read tasks, because the difference between internal and external production of the answers is not thought to be of much consequence.

Experiment 1

Method

Subjects. The subjects were 24 undergraduate students enrolled in an introductory psychology course at the University of Colorado, Boulder. Half of the subjects were tested during the Fall semester, and the other half were tested during the following Spring semester. They all received course credit for their participation.

Design and materials. All subjects served in each of four task conditions: read, calculate, generate, and verify. A 2 x 2 repeated measures design was used with two within-subjects variables. The first variable was locus of cognitive operations, with external locus (consisting of the read and calculate conditions) versus internal locus (consisting of the generate and verify conditions). The second variable was locus of stimulus production, with external locus (consisting of the read and verify conditions in which subjects saw the problem and the answer) versus internal locus (consisting of the generate and calculate conditions in which subjects saw the problem but had to produce the answer themselves). A preliminary analysis also included group (Fall subjects versus Spring subjects) as a between-subjects variable. Because there were no reliable differences between the Fall and Spring groups this factor was not included in the final analysis.

The stimulus materials consisted of index cards on which multiplication problems of the following type were written: 6 x 8 = 48 (for the read and verify conditions) and 6 x 8 = ? (for the calculate and generate conditions). There were five problems for each condition. The multiplication products consisted entirely of two-digit answers selected in the following manner.

There are 40 unique two digit multiplication products for the 2-times multiplication table through the 12-times multiplication table, excluding the answers 10 and 12. For answers having two or more possible pairs of multipliers (e.g., 6 x 8 = 48 and 4 x 12 = 48), one of the possible problems was randomly selected. From this reduced set of 40 problems, 20 problems were then randomly selected with the additional constraint that for any subset of answers beginning with the same digit, only half of the problems were selected (e.g., from 2 x 7 = 14, 3 x 5 = 15, 4 x 4 = 16, and 3 x 6 = 18, the two problems 3 x 6 = 18 and 3 x 5 = 15 were selected). Finally, the multipliers for half of the problems were

in ascending order (e.g., 9 x 11 instead of 11 x 9).

Procedure. All subjects performed the same 20 multiplication problems in 4 blocks of 5 problems each. The block orders and the problems within each block were the same for each subject. However, the task variable (read, calculate, generate, or verify) was counterbalanced using a Latin square procedure, so that for each group of four subjects, one subject was randomly assigned to one of four task orders. Thus, each subject participated in all four tasks, and across subjects the four tasks occurred equally often in each block. Before each presentation of a task block, the cards in that block were randomly reordered and placed before the subject. Subjects were then instructed to turn over a card every five seconds (prompted by the experimenter who used a digital watch for timing). The read and generate tasks were similar to those used in previous experiments (see Table 1 for an illustration of the four tasks). For the read condition, subjects simply read the problem (the multipliers) and the answer (the product) aloud. For the generate condition, subjects read the problem aloud then generated and said aloud the answer. For the calculate condition, subjects read the problem aloud and used a calculator to generate the answer, which they then read aloud. For the verify condition, subjects read the problem and the answer aloud and verified whether the answer was correct or not by saying "correct" or "incorrect." Because there were only five problems presented, none of the problems for the verify condition were actually incorrect, although subjects were told they would see problems that might be either correct or incorrect. The justification for this manipulation was that incorrect answers would have complicated the analysis and weakened the comparison with the other conditions, which necessarily included only correct answers. Further, given only five problems to verify, it seemed that this manipulation would still be effective in getting subjects to verify the answers.

Prior to receiving the multiplication problems, the subjects were informed that they would be given a retention test for the answers alone.

Insert Table 1 about here

After completing the 20 problems, subjects were given a 2 minute distractor task in which they named five associates for experimenter-supplied nouns. After the distractor task, subjects were asked to recall as many of the previous multiplication answers as possible and to write them down on an index card as they were recalled. Subjects had as much time as they needed to complete the task. Twelve of the subjects, three for each task order, were asked to think aloud while recalling the answers with minimal instructions for verbalization, as recommended by Ericsson and Simon (1984).

Results

The results are summarized in Table 2 in terms of proportions of correct recall responses for the four task conditions. Recall levels for the verify and generate conditions were higher than those for the read and calculate conditions. A repeated measures analysis of variance revealed a significant main effect of locus of cognitive operations, $F(1,23) = 47.82$, $MSe = .0391$, $p < .00001$. The main effect of locus of stimulus production was not statistically significant, $F(1,23) < 1$, nor was the interaction of the two factors, $F(1,23) < 1$. The results were as predicted: The tasks requiring internal cognitive operations showed a distinct retention advantage over the tasks that did not require such operations. Whether an answer was supplied by the experimenter or by the subject, however, did not reliably influence retention.

Table 1

Illustration of Sample Problem

Task	Subject Sees	Calculator Available	Subject Responds
Read	6 x 8 = 48	No	"6 x 8 = 48"
Generate	6 x 8 = ?	No	"6 x 8 = 48"
Calculate	6 x 8 = ?	Yes	"6 x 8 = 48"
Verify	6 x 8 = 48	No	"6 x 8 = 48, correct"

Insert Table 2 about here

The results concerning the verify condition argue against the concern raised in the Method section that including only correct answers to the multiplication problems might reduce the effectiveness of the request for the subjects to check the answers. Subjects' performance on the verify task was equal to that on the generate task and considerably greater than that on the read task, suggesting that subjects in the verify task did indeed check the answers as requested. If subjects had not actually checked the answers, the verify task would have been equivalent to a read task and hence performance on it would have been comparable to that on the read task. The results concerning the calculate condition have an interesting practical implication: Using a calculator to perform arithmetic computations may lead to reduced retention of the computed answers. More generally, the results from all four conditions taken together have an important theoretical implication: The typical generation effect is due to internal cognitive operations rather than internal stimulus production or generation per se.

Experiment 2

Because of the important practical and theoretical implications of the results in Experiment 1, we aimed in Experiment 2 to assess their generalizability. More specifically, our goal was to replicate and extend the findings from Experiment 1 along two dimensions. First, we sought to determine whether the same pattern of results would obtain for retention over considerably longer delays than were involved in the immediate testing situation of Experiment 1. The earlier studies of the generation effect (e.g., Slamecka & Graf, 1978) were limited almost exclusively to short retention intervals. Would

Table 2

Proportion of Answers Correctly Recalled in Experiment 1 as a Function of Locus of Cognitive Operations (Internal or External) and Locus of Stimulus Production (Internal or External)

	Cognitive Operations	
Stimulus Production	Internal	External
	Verify	Read
External	.68	.38
	Generate	Calculate
Internal	.68	.42

the generation effect and the advantage we found for internal cognitive operations be obtained even for retention intervals as long as one week? Second, we aimed to assess whether a recognition test procedure would lead to the same findings as the recall procedure used in Experiment 1. This second question is related to the first because it seems likely that recall performance would be close to the floor after a long delay, whereas a recognition test might prove to be more sensitive. Both recall and recognition tests have been used in some previous investigations of the generation effect (again see, e.g., Slamecka & Graf, 1978).

Method

Subjects. The subjects were 48 undergraduate students from the same population employed in Experiment 1 but tested the following Fall semester. None of the subjects in Experiment 1 also participated in Experiment 2. Again subjects received credit in an introductory psychology course for their participation. The subjects were divided into three conditions with 16 subjects in each condition. The assignment of subjects to conditions was determined on the basis of time of arrival for testing.

Design and materials. The design of the experiment was the same as that of Experiment 1 with two modifications. First, a between-subject variable was added: retention interval condition. Subjects were tested either immediately, after a two-day delay, or after a seven-day delay. Second, two different measures of retention were employed: recall and recognition.

The same materials, consisting of multiplication problems written on index cards, were used in the study phase of this experiment as had been used in Experiment 1. Specifically, 20 multiplication problems were randomly selected from a set of 40, each of which had a different two-digit answer and both

multipliers within the range from 2 to 12, and half of which had the multipliers in ascending order. The remaining 20 problems from the set of 40 were used as distractors for the recognition test. Each of the products from the 20 distractors was randomly paired with a different product from one of the 20 study problems to form the forced choice recognition test. The order of the two products in a pair was random with the constraint that for 10 of the 20 pairs, the study product preceded the distractor product. The recognition test was written on a single sheet of paper, with the product pairs written in two vertical columns of 10 pairs each.

A 521-word prose passage was employed in a filler task for the two-day and seven-day retention interval conditions. The passage was typed with single-spacing on a single sheet of paper.

Procedure. The procedures for the study phase of the experiment were identical to those used in Experiment 1, except that the subjects were not warned that they would be given a retention test. This modification was made because we did not want to encourage the subjects in the two-day and seven-day retention interval conditions to rehearse the problems during the long delay before the retention test. Also, previous studies (e.g., Slamecka & Graf, 1978) have demonstrated that the generation effect occurs under either intentional or incidental instructions.

The same distractor task was used as in Experiment 1 except that subjects were given only 1.5 rather than 2 minutes to complete that task. The recall task was also the same as in Experiment 1. Immediately following the recall task subjects were given the recognition test. For each pair of products the subjects were to circle the one number in the pair that was an answer to one of the multiplication problems they were given during the study phase.

For subjects in the immediate retention interval condition, as for subjects in Experiment 1, the retention tests were administered immediately after the distractor task. Instead of the retention tests, the subjects in the two-day and seven-day retention interval conditions were given the materials for the filler task. Specifically, they were given a sheet of paper containing the prose passage and told to take the passage home and read it sometime before they returned for their scheduled second session. They were told that while reading the passage, they should circle all of the *t*'s in it. Further, they were told to read the passage only once and after finishing it to write the time and date on the paper then to return the paper at the second session. This filler task was given simply to discourage subjects from rehearsing the multiplication problems during the retention interval.

Results

Recall. The results of the recall task are summarized in Table 3 in terms of proportions of correct recall responses for the four tasks in each of the three retention interval conditions. As in Experiment 1, recall levels for the generate and verify conditions were higher than those for the read and calculate conditions, and this same pattern of results was found for each of the three retention interval conditions although increased delay between study and test did depress performance levels considerably. A mixed analysis of variance revealed a significant main effect of locus of cognitive operations, $F(1,45) = 46.05$, $MSe = .0462$, $p < .00001$, but the main effect of locus of stimulus production was not statistically reliable, $F(1,45) < 1$, although the interaction of these two factors did approach statistical significance, $F(1,45) = 3.47$, $MSe = .0438$, $p < .10$, presumably because the generate task yielded somewhat better performance overall than the verify task but the read task yielded somewhat better performance overall than the calculate task. The main effect of

retention interval condition was significant, $F(2,45) = 16.53$, $MSe = .0644$, $p < .0001$, but this factor did not enter into any significant interactions.

Insert Table 3 about here

Recognition. The results of the forced-choice recognition task are summarized in Table 4 in terms of proportions of correct recognition responses for the four tasks in each of the three retention interval conditions. Although performance levels for the recognition task were higher than for the recall task, the same pattern of results was found for recognition as for recall. Specifically, recognition levels were higher for shorter delays between study and test and, most crucially, were higher for the generate and verify conditions than for the read and calculate conditions, with essentially no differences between the generate and verify or between the read and calculate conditions. A mixed analysis of variance yielded reliable main effects of retention interval condition, $F(2,45) = 6.02$, $MSe = .0464$, $p < .01$, and of locus of cognitive operations, $F(1,45) = 11.69$, $MSe = .0500$, $p < .01$, but not of locus of stimulus production, $F(1,45) < 1$. The only interaction that approached statistical significance was that involving retention interval condition and locus of stimulus production, $F(2,45) = 2.44$, $MSe = .0298$, $p < .10$.

Insert Table 4 about here

Experiment 3

The standard generate task differs from the standard read task along two dimensions. First, the subject must supply the stimulus in the generate task, whereas the experimenter supplies the stimulus in the read task. Second, the

Table 3

Proportion of Answers Correctly Recalled in Experiment 2 as a Function of Locus of Cognitive Operations (Internal or External), Locus of Stimulus Production (Internal or External), and Retention Interval Condition (Immediate, Two-day, Seven-day)

Stimulus Production and Retention Interval	Cognitive Operations: Internal	Cognitive Operations: External
External	Verify	Read
Immediate	.59	.42
Two-day	.40	.24
Seven-day	.24	.10
Mean	.41	.25
Internal	Generate	Calculate
Immediate	.55	.34
Two-day	.49	.16
Seven-day	.40	.14
Mean	.48	.21

Table 4

Proportion of Answers Correctly Recognized in Experiment 2 as a Function of Locus of Cognitive Operations (Internal or External), Locus of Stimulus Production (Internal or External), and Retention Interval Condition (Immediate, Two-day, Seven-day)

Stimulus Production and Retention Interval	Cognitive Operations: Internal	Cognitive Operations: External
External	Verify	Read
Immediate	.82	.81
Two-day	.76	.61
Seven-day	.72	.52
Mean	.77	.65
Internal	Generate	Calculate
Immediate	.81	.65
Two-day	.75	.66
Seven-day	.69	.64
Mean	.75	.65

subject must perform the relevant cognitive operations in the generate task, whereas the experimenter or another agent performs those operations in the read task. In Experiments 1 and 2 we sought to determine which of these dimensions was crucial for the generation effect. The answer we obtained was clear-cut: The locus of stimulus production had essentially no effect, whereas the locus of cognitive operations had a major effect. That is, whether the subject or the experimenter supplied the answers to multiplication problems proved to be immaterial, but whether the subject or some other agent performed the relevant multiplication operations greatly affected the subject's memory for the answers to the problems. This pattern of results was found in Experiment 1 for an immediate recall test and in Experiment 2 for both recall and recognition tests conducted immediately, after a two-day delay, and after a one-week delay.

These findings provide support for a cognitive operations hypothesis which seems most closely aligned with the general proceduralist account proposed by Kolers and Roediger (1984) and the more specific account proposed by Glisky and Rabinowitz (1985) but is also consistent with both types of explanations that have been proposed to account for the generation effect, those concerning effort and those concerning lexical activation. By both types of explanations, the crucial aspect of generation is the inducement of auxiliary mental processes or cognitive operations. In the present experiments the relevant mental processes were the multiplication operations. In the generate task the subjects had to perform those operations in order to derive the answers to the multiplication problems. Although the answers to the problems were provided in the verify task, subjects had to perform the multiplication operations in order to verify that the answers were correct. In contrast, no multiplication was necessary in the read task because the experimenter supplied the answers and the subjects were not requested to check them. The answers were not supplied by the

experimenter in the calculate conditions, but a calculator rather than the subjects themselves performed the multiplication in that case.

It could be argued that some cognitive operations must be performed by subjects in the calculate condition even though the multiplication operations were not necessary. For example, the subjects had to decide which calculator buttons to press and the order in which to press them. However, these operations may not be relevant for two related reasons. First, because no calculator was present during retention, the subjects could not readily reactivate at test the cognitive operations used at study to derive the answers with the calculator. Second, the button press operations may be so similar for all the multiplication problems that even if subjects are reminded of these operations at test, they cannot use this information to differentiate the study problems from others like them. At the basis of both of these reasons is the assumption that cognitive operations performed at study may be useful only if such operations can be employed at test as successful retrieval cues (see also Kolers & Roediger, 1984). In fact, in the verbal protocols we collected in Experiment 1 during recall, subjects frequently recalled the two numbers multiplied together and used these numbers as cues to remind themselves of the multiplication product.

It would thus seem inappropriate to explain the generation effect simply in terms of the number of cognitive operations performed. Rather, it seems preferable to focus on the type of cognitive operations performed and assess to what degree a specific type of operation aids retrieval. Some experimenters have attempted just this task.

Nairne and Widner (1987), for example, have performed a series of experiments demonstrating a generation effect for nonwords, provided an appropriate test is used to measure retention. Nairne and Widner have

hypothesized that when subjects generate a fragment to complete a nonword, the unit of generation is the fragment itself, whereas when subjects generate a fragment to complete a word, the unit generated is the whole word. Thus the lack of a generation effect for nonwords occurs because the usual retention test that requires subjects to recognize or recall whole nonwords is not testing the unit that subjects actually generated. By testing retention for the fragment subjects had generated, Nairne and Widner were able to obtain a generation effect for the nonword fragments. Furthermore, by having subjects regenerate the nonwords at test using the same procedure used during generation, then testing subjects with a recognition procedure, a generation effect was obtained for the nonwords themselves. These results emphasize the importance of determining exactly what is generated in a given generate task and using an appropriate test to assess what is retained.

Other experiments support this idea of test appropriateness: that the effectiveness of generation for retention depends upon how items are generated and the types of information present at retrieval. For example, Rabinowitz and Craik (1986) have found that if words are generated using associative cues, a generation effect is obtained when the same associative cues are present at retrieval but not when the cues tap qualitatively different types of information (e.g., extralist rhyme cues). The reverse is also the case: Words generated using rhyme cues show a generation effect when rhyme cues are present at retrieval but not when associative cues are present. Similarly, Glisky and Rabinowitz (1985) found greater recognition performance when the same generation operations were present at study and test (filling in the same two missing letters of a word) than when different generation operations were involved (filling in a different pair of letters at study and test).

Experiment 3 was designed to assess the influence of test appropriateness in the present paradigm and to gain further insight into why the calculate condition yielded poorer retention than the generate condition. Towards these ends, the generate and calculate study conditions were crossed with two comparable test conditions. That is, subjects either generated or calculated the answers to multiplication problems in the study phase and then either generated or calculated the answers to the same problems intermingled with distractor problems in the test phase. The retention test consisted of a recognition rating by which subjects indicated how certain they were that a given problem performed in the test phase was one that they had performed in the study phase. If test appropriateness is the major determinant of retention, then we should find an interaction of study condition and test condition such that subjects are able to recognize a problem better when the study and test conditions match than when they differ. Alternatively, if the mental operations performed in the calculate condition cannot be used as successful retrieval cues even when the subjects are reminded of these operations at test, then, as in Experiments 1 and 2, we should find a main effect of study condition such that subjects are able to recognize a problem better when they generate the answer to the problem during the study phase than when they use a calculator to produce the answer at study.

Method

Subjects. The subjects were 24 undergraduates from the same population used in Experiments 1 and 2 but tested the following Spring semester. None of the subjects in Experiment 3 had participated in Experiments 1 or 2. As previously, subjects received credit in an introductory psychology course for their participation. The subjects were divided into four groups with six subjects in each group. These groups were used to counterbalance the order of

conditions (see the section on design and materials). The assignment of subjects to groups was determined on the basis of time of arrival for testing.

Design and materials. Only two of the study conditions used in Experiments 1 and 2 were employed in this experiment: generate and calculate. These study conditions were crossed with two analogous test conditions: generate and calculate. All subjects were exposed to all conditions; hence, a 2 x 2 repeated measures design was used, with the variables study condition and test condition.

The same 40 multiplication problems were used in this experiment as had been used in Experiment 2. As in Experiment 2, 20 of these problems were used in the study phase, and all 40 were used in the test phase. The 20 study problems were divided into two 10-problem blocks, and the 40 test problems were divided into two 20-problem blocks. Each test block included five problems from each of the two study blocks intermixed with ten distractors. The study problems were presented in a fixed pseudorandom order with the constraint that successive problems not include any of the same multipliers. Similarly, the test problems were presented in a fixed pseudorandom order with the constraint that no more than two distractor problems or two problems from the same study block occur successively. The block orders and the order of problems within each block were the same for each subject. However, the study task and the test task were counterbalanced across subjects. Four subject groups were employed to counterbalance condition order. For the first group, the generate task preceded the calculate task at study and at test; for the second group, the generate task preceded the calculate task at study but the calculate task preceded the generate task at test; for the third group, the calculate task preceded the generate task at study but the generate task preceded the calculate task at test; and for the fourth group, the calculate task preceded the generate task at both study and test.

The multiplication problems were typed on sheets of paper, one problem per line. A cover sheet containing a slit in the center was used to allow the subjects to view one problem at a time.

Procedure. A self-paced procedure was used to display the multiplication problems to each subject. The subjects used the cover sheet to allow one problem to be displayed at a time. After completing the problem, they immediately advanced to the next problem by sliding the cover sheet down the page. Otherwise, the procedures for the study phase of the experiment were the same as those used in the generate and calculate conditions of Experiments 1 and 2. As in Experiment 2, subjects were not warned that they would be given a recognition test. The same distractor task was used as in Experiments 1 and 2, with a two-minute duration, as in Experiment 1. The test phase followed immediately after the distractor task.

Unlike Experiments 1 and 2, the test phase involved performing multiplication problems as well as recognition for the problems performed during the study phase. For the generate test condition, subjects read the problem aloud, then generated and said aloud the answer. For the calculate test condition, subjects read the problem aloud as they entered it into the calculator to compute the answer, which they then read aloud from the calculator display. In each test condition, after saying the answer to the problem, the subjects were to indicate whether the problem was "old" (performed earlier in the study phase) or "new" (not performed earlier) and to give a recognition rating on a scale from 1 to 6, with 1 meaning they were sure that the problem was new and 6 meaning they were sure that the problem was old. After saying aloud the recognition rating, the subject advanced to the next problem, using the same cover sheet mechanism employed in the study phase.

Results

The results are summarized in Table 5 in terms of mean recognition ratings on the six-point scale for the old and new test items as a function of the locus of cognitive operations at test. A repeated measures analysis of variance revealed a significant effect of item type, $F(1,23) = 101.62$, $MSe = 0.8733$, $p < .00001$, indicating that subjects were able to discriminate old items ($M = 4.506$) from new items ($M = 2.583$). The locus of cognitive operations at test (external in the calculate condition and internal in the generate condition) did not significantly influence either the overall recognition ratings, $F(1,23) < 1$, or the difference in ratings between old and new items, $F(1,23) = 2.55$, $MSe = 0.1830$, $p > .10$.

Insert Table 5 about here

Most crucial to the present concerns are the combined effects on the recognition ratings for the old items of the locus of cognitive operations at study and the locus of cognitive operations at test. These results are summarized in Table 6. As in Experiments 1 and 2, the locus of cognitive operations at study had a profound impact on memory; the recognition ratings for items generated in the study phase ($M = 4.850$) were considerably higher than those for items calculated in the study phase ($M = 4.163$), $F(1,23) = 15.51$, $MSe = 0.7316$, $p < .001$. In contrast, the locus of cognitive operations at test did not have a significant main effect on the recognition ratings, $F(1,23) = 1.64$, $MSe = 0.4274$, $p > .10$, and the interaction of the locus of cognitive operations at test and at study was only marginally significant, $F(1,23) = 3.12$, $MSe = 0.3204$, $p < .10$. The pattern of means was in the direction of test appropriateness (the ratings tended to be higher when the test condition matched

Table 5

Mean Recognition Rating (Six-Point Scale) on Old and New Items in Experiment 3 as a Function of Locus of Cognitive Operations (Internal or External) at Test

Cognitive Operations at Test	Item Type Old	New
External (Calculate)	4.617	2.554
Internal (Generate)	4.396	2.612

the study condition than when it differed from the study condition), but the effect of test appropriateness was overwhelmed by the effect of study condition.

Insert Table 6 about here

General Discussion

In three experiments support was provided for an explanation of the generation effect in terms of promoting cognitive operations at the time of study that can be used as successful retrieval cues at the time of retention testing. Experiments 1 and 2 compared the importance of the locus of cognitive operations at study to the locus of stimulus production at study. Whereas the locus of stimulus production (whether the subject or the experimenter supplied the to-be-remembered stimulus) had essentially no effect on retention, the locus of cognitive operations (whether the subject or another agent performed the relevant cognitive operations) had a major effect. Further, Experiment 3 compared the importance of the locus of cognitive operations at study to the locus of cognitive operations at test. Although there was a trend indicating the influence of test appropriateness, or the match between cognitive operations at study and at test, the locus of cognitive operations at study had a much greater influence than the locus of cognitive operations at test. More specifically, subjects in the present experiments performed simple multiplication problems. In Experiments 1 and 2, subjects showed superior retention for the answers to these problems when they performed the multiplication operations themselves, but their retention was not influenced by whether they supplied the answers to the problems. Likewise, in Experiment 3, subjects showed better memory for multiplication problems when they performed the multiplication themselves at the time of study, but their retention was

Table 6

Mean Recognition Rating (Six-Point Scale) on Old Items in Experiment 3 as a Function of Locus of Cognitive Operations (Internal or External) at Study, and Locus of Cognitive Operations (Internal or External) at Test

Cognitive Operations at Test	Cognitive Operations at Study: Internal	Cognitive Operations at Study: External
External	Generate/Calculate 4.833	Calculate/Calculate 4.350
Internal	Generate/Generate 4.867	Calculate/Generate 3.975

influenced much less by whether they performed the multiplication themselves at the time of test.

One result of previous studies has been that generate rules regardless of their triviality result in superior retention. For example, generation effects have been found even when subjects only have to transpose two letters in order to generate the stimulus word (e.g., Nairne & Widner, 1987) or they have to supply only a single letter and that letter is always the same (e.g., Donaldson & Bass, 1980). The present study indicates, however, that not all forms of generation are sufficient to yield this retention advantage. If the subject uses an external device (i.e., a calculator) for generation, no retention advantage is found.

In each of the three present experiments, we found that using a calculator to perform the multiplication operations when studying the problems was less effective in promoting retention than mentally performing the multiplication operations without any external aid. We attribute this benefit of mental multiplication operations to the fact that these operations can act as retrieval cues at the time of retention testing. The mental multiplication operations for different problems are well differentiated so that if the subjects can remember which operations were performed they may be reminded of the specific problem and answer. In contrast, the mental operations involved in using the calculator to perform multiplication are probably not well differentiated but rather may involve very similar button presses for the different problems. The small differences in the patterns of button press operations across problems may be responsible for the marginal effect of test appropriateness that was observed in Experiment 3.

The retention advantage for performing multiplication operations mentally, rather than with the aid of an external device, has an important practical

implication, as mentioned in the Results section of Experiment 1. For situations in which it is crucial to remember the answer to a multiplication problem, it is better to perform the calculation in one's head rather than with a calculator. A second practical implication follows more indirectly from the present results. In the present situation adult subjects were employed who were well trained at multiplication. Perhaps, however, a similar pattern of results would be obtained for children learning the multiplication table. Retention of the answers to multiplication problems may also be hindered in that case when a calculator is employed. Hence, from an educational standpoint, it may be appropriate to discourage the use of calculators by students, who may otherwise have difficulty acquiring the necessary multiplication facts.

The three present experiments were limited exclusively to situations involving multiplication problems. Hence, we have no guarantees that the conclusions we reached can be generalized to other situations in which the generation effect has been found that do not involve arithmetic operations. Nevertheless, we have shown that these effects with multiplication hold under a variety of conditions--with recall or recognition testing, under intentional or incidental learning instructions, at retention intervals varying from immediate to seven days, and when only the answer must be remembered or when the problem as a whole must be retained. Hence, we are confident that the locus of multiplication operations is an important factor in influencing memory. Future research should be aimed at assessing the extent to which other cognitive operations have similar effects on memory. In other words, we propose that future experiments to understand the generation effect should focus on identifying the specific mental operations involved in different generation tasks as well as the relationship between information and mental operations present at generation and at retrieval.

References

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. Memory & Cognition, 6, 20-25.

Banaji, M. R., & Greenwald, A. G. (1984, May). Self-generated information aids memory at retrieval. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, Ill.

Donaldson, W., & Bass, M. (1980). Relational information and memory for problem solutions. Journal of Verbal Learning and Verbal Behavior, 19, 26-35.

Ellis, H. C., Thomas, R. L., & Rodriguez, I. A. (1984). Emotional mood states and memory: Elaborative encoding, semantic processing, and cognitive effort. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 470-482.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, Massachusetts: The MIT Press.

Gardiner, J. M., & Rowley, J. M. C. (1984). A generation effect with numbers rather than words. Memory & Cognition, 12, 443-445.

Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 732-741.

Glisky, E. L., & Rabinowitz, J. C. (1985). Enhancing the generation effect through repetition of operations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 193-205.

Graf, P. (1980). Two consequences of generating: Increased inter- and intraword organization of sentences. Journal of Verbal Learning and Verbal Behavior, 19, 316-327.

Greenwald, A. G. (1981). Self and memory. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 15, pp. 201-236). New York: Academic Press.

Griffith, D. (1976). The attentional demands of mnemonic control processes. Memory & Cognition, 4, 103-108.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. Journal of Verbal Learning and Verbal Behavior, 17, 649-667.

Jacoby, L. J. (1983). Remembering the data: Analyzing interactive processes in reading. Journal of Verbal Learning and Verbal Behavior, 22, 485-508.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. Journal of Verbal Learning and Verbal Behavior, 23, 425-449.

McDaniel, M. A. (1981). Syntactic complexity and elaborative processing. Memory & Cognition, 9, 487-495.

McElroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect. Journal of Verbal Learning and Verbal Behavior, 21, 249-259.

McFarland, C. E., Jr., Frey, T. J., & Rhodes, D. D. (1980). Retrieval of internally versus externally generated words in episodic memory. Journal of Verbal Learning and Verbal Behavior, 19, 210-255.

Nairne, J. S., Pusen, C., & Widner, R. L., Jr. (1985). Representation in the mental lexicon: Implications for theories of the generation effect. Memory & Cognition, 13, 183-191.

Nairne, J. S., & Widner, R. L., Jr. (1987). Generation effects with nonwords: The role of test appropriateness. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 164-171.

Payne, D. G., Neely, J. H., & Burns, D. J. (1986). The generation effect: Further tests of the lexical activation hypothesis. _Memory & Cognition_, _14_, 246-252.

Rabinowitz, J. C., & Craik, F. I. M. (1986). Specific enhancement effects associated with word generation. _Journal of Memory and Language_, _25_, 226-237.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. _Journal of Experimental Psychology: Human Learning and Memory_, _4_, 592-604.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. _Journal of Verbal Learning and Verbal Behavior_, _22_, 153-163.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. _Journal of Experimental Psychology: Human Learning and Memory_, _5_, 607-617.

Zacks, R. T., Hasher, L., Sanft, H., & Rose, K. C. (1983). Encoding effort and recall: A cautionary note. _Journal of Experimental Psychology: Learning, Memory, and Cognition_, _9_, 747-756.

Author Notes

This research was supported in part by United States Air Force Human Resources Laboratory Contract Number VE5744-022-001 and United States Army Research Institute Contract Number MDA903-86-K-0155. We wish to thank Lyle Bourne, Anders Ericsson, and Jim Nairne for helpful discussions about this research, Martha Neal for help preparing the materials, testing subjects, and scoring the response protocols in Experiment 2, Laurelyn Andrews for help preparing the materials, testing subjects, scoring the response protocols, and analyzing the data in Experiment 3, and Antoinette Gesi for help scoring the response protocols and analyzing the data in Experiments 2 and 3. Parts of this study concerning Experiments 1 and 2 were reported at the Symposium on Recent Advances in Experimental Cognitive Psychology at the Rocky Mountain Psychological Association Meeting in Albuquerque, New Mexico on May 1, 1987. Requests for reprints should be sent to either Robert J. Crutcher or Alice F. Healy, Department of Psychology, University of Colorado, Campus Box 345, Boulder, CO 80309.

APPENDIX E

LONG-TERM RETENTION OF ALGEBRA

Lori Meiskey
Alice F. Healy
Robert W. Ellingwood
Lyle E. Bourne, Jr.
University of Colorado

Educators hope that the knowledge and skills which students have acquired will be long-lasting. Few studies have been conducted, however, which address the issue of the long-term retention of knowledge and skills, particularly of knowledge and skills learned in a naturalistic (i.e., versus laboratory) situation. The most notable exception to this dearth of studies is the work of Harry Bahrick. Bahrick has assessed the long-term (in some studies up to 50 years after the original acquisition) retention of knowledge (e.g., memory for the names and faces of high school colleagues, Bahrick et al., 1975; memory for locations and names of streets in a college town, Bahrick, 1983; memory for Spanish, Bahrick, 1984). The current work is strongly based on Bahrick's methodology, particularly Bahrick's (1984) study of the long-term retention of Spanish learned in school. But this work attempts to extend as well as utilize Bahrick's findings and methdology in studying the long-term retention of knowledge and skills.

The domain of study was College Algebra. This domain was chosen for several reasons. One reason was that its use should be easily identifiable by students. That is, at some period of time after the students had completed the course, they should easily remember how much they had used algebra since the completion of the course. Another reason was that there was a sufficient literature available upon which a test of algebra could be constructed which would take account of students' knowledge states of and skill level at using algebra (e.g., Matz, 1982; Carry, Lewis, & Bernard, no date; Sleeman, 1982, 1984). Also, students taking the algebra course were likely to be freshman, and it was hoped that they could be easily contacted throughout the next three years of their college career to return for retesting.

Bahrick (1984) used two basic sources of information in examining the long-term retention of Spanish learned in school. One source was a test given

to subjects which assessed various components of Spanish knowledge (e.g., vocabulary, grammmar). The other source of information was a questionnaire which assessed the original level of acquisition of Spanish (e.g., number of years taken of Spanish, course grade), and also assessed the maintenance of that knowledge in the interval since the subjects had taken their last Spanish course (e.g., whether Spanish is spoken in the home). The present study also utilizes these two sources of information. A test was designed to include various components of algebra. A questionnaire was designed which assessed the original acquistion level of algebra (e.g., how many algebra courses the student had taken before the target algebra course) and the maintenance of that knowledge (e.g., what courses did the student take that involved the use of algebra in the interim between the end of the course and the retest). Information about the level of acquisition was also available from a test given at the end of the course. Bahrick used a multiple regression technique to analyze the factors important in predicting retention level, and this method is also borrowed in the current study. This study differs from that of Bahrick (1984) in that it is longitudinal, whereas his Spanish study was cross-sectional.

The purpose of the present study is to examine the types of knowledge and skills which are lost or retained after some period of disuse. An attempt was also made to assess the effects of maintenance, or use of algebra during the retention interval, on retention of algebra.

Study 1

Method

Subjects

Subjects were students enrolled in a lower division algebra course at the University of Colorado at Boulder. All entering freshman are required to take

a mathematics test by the university. The students in the algebra course being tested were placed in that course due to low scores (but not scores at the bottom of the curve) on this university test. Eighty-six students in one class were tested near the end of their course in the Spring semester of 1987. The score on this test was counted toward the student's final grade. Fifteen of these subjects returned approximately six months later for retesting. They were each paid seven dollars for approximately one hour and fifteen minutes of testing.

Materials

Two versions of a 60-item multiple choice algebra test were constructed. The second version of the test contained the same items as the first version. However, the variable labels differed between tests. For example, a question in the first version of the test was: "x/2 + y/3 is equal to:"; the comparable item from the second version of the test was: "m/2 + n/3 is equal to:". Also, the second version of the test contained a different random ordering of both items and of choices within an item from the first test version.

The categories of items to be tested were determined by an analysis of the textbook which the students used (Swokowski, 1986). The choice of categories was further constrained by a decision to test only information which pertained to the use of algebra in manipulating equations. Therefore, categories of potential items such as word algebra problems and graphing of equations were not considered. The categories of items to be tested were intended to capture the knowledge and skills which would be necessary in order for a student to manipulate and understand equations successfully. This encompassed most of the first two chapters of the textbook. The categories determined using the algebra textbook as well as the above constraint are listed in Table 1 with an example question from each category.

Table 1
Categories of items used in the algebra test with example questions from the test.

1. Use of the quadratic formula.
   The equation $2x^2 - 8x + 3 = 0$ has: [number of roots]
2. Complete the square.
   Complete the square for $x^2 + 7x$.
3. Combining exponents of common terms.
   What does $2^m 2^n$ equal?
4. Manipulating equations.
   a. Getting a common denominator.
      $x/2 + y/3$ is equal to:
   b. Simplifying an expression.
      Simplify $3z^2 - 18z + 11 = 3z + 5$
   c. Order of operations.
      Let $z = 2$; Solve for y: $(4/z+5*z)/y = 1$
   d. Multiple-term products.
      $(2p + 3)(p + 5)$ equals:
   e. Cross-multiplication.
      If $2/b = c/7$, then what is c equal to?
5. Exponentiation.
   $(3a)^3$ is equal to:
6. Absolute value.
   If $t = 5$, then $|t - 12|$ equals:
7. Factoring Equations.
   Factor $x^2 + 3x + 2 = 0$
8. Properties and laws.
   a. Theorem of zero.
      When does $z*0 = z$?
   b. Negative numbers.
      $-(-a)$ is positive only when:
   c. Trichotomy.
      If $x > 0$, then x could possibly be equal to:
   d. Law of signs.
      If a and b have the same sign, or one or both of them equal zero, then ab is: [less than, greater than, or equal to zero]
   e. Substitution.
      If $ab = 10$ and $c = b$, then ac equals:
   f. Commutative property.
      If $m + 5 = 5 + n$, then n is equal to:
   g. Associative property.
      Is $3(cd)$ equal to $(3c)d$?
   h. Distributive property.
      What is the final simplified form of $6 + (3 + a)2$?
9. Square roots.
   If $a = 2$ and $b = 6$, then $\sqrt{a + b}$ equals:
10. Distractor problem (answer was "none of the above").
    $8p + 10q = 5$ is equivalent to:

Questions which would test these categories of knowledge were constructed from a number of sources. The first source was the textbook itself. Other sources of questions were a task analysis of algebra (Bundy, 1975) and experiments which examined the types of errors that students make in solving and manipulating equations (e.g., Matz, 1982; Carry et al., no date; Sleeman, 1982, 1984). The last source of items were those created by the authors. The alternative choices for each item were formed using the last two sources. In particular, the work of Matz, Carry et al., and Sleeman were instrumental in creating alternative choices which embodied plausible errors that students might make. The final approval of the suitability of the test was determined by the course instructor (R.W.E.).

A questionnaire based loosely on Bahrick's (1984) study of the long-term retention of Spanish was developed to examine the students' previous knowledge of algebra, and their use of algebra subsequent to the termination of the algebra course. The questionnaire contained a number of rating scales on which the student would indicate the frequency of use of algebra in particular situations (e.g., converting temperature from Fahrenheit to Centigrade or vice-versa).

The questionnaire also contained two tables in which the student was to fill out the relevant information for courses taken in middle or junior high school, in high school, and in college. The tables contained columns for the date the course was completed, the student's grade level when taking the course (e.g., ninth grade, junior in college), the number of semesters of this course completed, and the final grade obtained for the course. The tables also contained columns in which the subject would indicate whether they had used arithmetic, fractions, and/or equations in the course. The first table listed possible mathematics courses taken in middle school through college (e.g.,

Algebra, Geometry, Calculus). The second table listed possible science courses in which algebra might be used (e.g., Physics, Biology, Chemistry). Also listed in each table was a category labelled "Other" in which the student could fill in courses which were not already listed.

<u>Procedure</u>

The students were tested at the end of the semester, but before the final test was given for the course. Half of the students were given the first version of the test, and the other half of the students were given the second version of the test. Before starting the test, the students were requested to fill out a form to indicate their current address and phone numbers. They were informed that they would be contacted at a later date to return for testing. They were also told that if they chose to be retested, they would receive payment for their time. The students were given 50 minutes to complete the test. Testing occurred during one of the students' regular class periods.

Fifteen of these subjects returned for testing approximately 6 months after the end of the Spring 1987 semester. The students were given the alternative version of the test from the one they had taken at the end of the semester. For example, if a student took version one of the test at the end of the semester, that student took version two of the test at the time of retesting. After completing the retest, students completed the questionnaire. Students were given one hour and fifteen minutes to complete both the test and the questionnaire.

Results

Fifteen subjects who had taken the algebra test at the end of the Spring 1987 semester returned for retesting between 5 months and 6.8 months after they had completed the end of the Spring 1987 semester (Mean = 6.0). They showed no significant change in retest percent correct (Mean=81.5) from their

scores on the end-of-semester test (Mean=80.3), ($t=0.62$, $p>.05$). However, an analysis of individual items revealed specific losses and items with no loss from the end-of-semester test to the retest.

Three categories of items showed a loss in percent correct from the end-of-semester test to the retest. See Table 2 for the frequencies of items by category which showed a loss or no loss (i.e., a gain or no change) in percent correct on the retest. Also see Table 3 for the mean percent correct for each category of items for both the end-of-semester test and the retest. For the categories 'use of the quadratic formula,' 'completing the square,' and 'combining exponents of common terms,' all items declined in percent correct on the retest. Other categories, however, showed no loss on the retention test, including all but one of the categories listed under 'manipulating equations' ('finding a common denominator' was the exception), 'exponentiation,' and 'absolute value' (also note that there was one exception within each of the categories of 'products' and 'exponentiation'). For other categories of items, the pattern of loss or no loss was not clear.

A stepwise regression was performed with retest score as the dependent variable and end-of-semester test score as an independent variable. Other independent variables were derived from the questionnaire: final course grade, whether or not the student was currently taking a mathematics course involving the use of algebra, a score based on the frequency of algebra use since the end of the algebra course, and the number of previous courses taken in mathematics which involved using algebra. Final grade was the strongest predictor of retest score. Final grade predicted the greatest amount of variance, and once this was accounted for, no other independent variables were significant predictors of retest score (for Final grade alone as a predictor of retest score, $F=29.6$, $p<.001$). End-of-semester test score, when entered

Table 2
Number of items in each category which exhibited a loss or no loss in percent correct from the end-of-semester test to the retest.

Category	# Items with Loss	# Items with No Loss
Quadratic formula	3 ***	0
Complete the square	2 *	0
Combine exponents of common terms	3 *	0
Manipulating Equations		
Common denominator	4 *	3 **
Simplify expression	0	3 **
Order of operations	0	2 *
Products	1	3
Cross-multiplication	0	2
Exponentiation	1	4
Absolute value	0	2
Factoring equations	1	2
Properties & Laws		
Theorem of Zero	1 *	1
Negative numbers	1 *	1 *
Trichotomy	1	1
Law of Signs	1	1
Substitution	2	1
Commutative	1	3
Associative	2	2
Distributive	0	2
Square roots	1	1
Distractor	1	0
Total	26	34

* Number of asterisks indicates the number of items on which <50% of the students who took the pretest in Study 2 were correct on the pretest.

Table 3
Percent correct for items averaged over categories for Study 1 students (N=15) end-of-semester test and retest, and for Study 2 students (N=75) pretest.

Category	Number of Items	Study 1 End-of-semester Mean %Correct	Study 1 Retest Mean %Correct	Study 2 Pretest Mean %Correct
Quadratic formula	3	55.3	44.7	29.7
Complete the square	2	73.0	63.5	47.0
Combine exponents of common terms	3	82.3	60.0	63.3
Manipulating Equations				
Common denominator	7	76.0	80.1	57.3
Simplify expression	3	51.3	62.3	44.7
Order of operations	2	56.5	66.5	59.5
Products	4	88.5	93.3	92.8
Cross-multiplication	2	83.0	96.5	83.5
Exponentiation	5	88.0	90.6	78.8
Absolute value	2	83.5	93.5	90.0
Factoring equations	3	88.7	89.0	82.0
Properties & Laws				
Theorem of Zero	2	76.5	60.0	74.0
Negative numbers	2	30.0	33.5	35.0
Trichotomy	2	96.5	90.0	97.5
Law of Signs	2	100.0	95.0	92.0
Substitution	3	100.0	95.0	90.8
Commutative	4	95.0	94.8	92.3
Associative	4	91.5	85.0	87.0
Distributive	2	96.5	100.0	90.0
Square roots	2	93.5	90.0	93.0
Distractor	1	73.0	60.0	68.0

into the equation first, was predictive of retest score ($F=13.7, p<.01$); however, this score was highly correlated with Final grade ($r=0.82$), and so both variables predicted essentially the same portion of variance. No other correlations between variables were significant.

There were no significant differences in percent correct between test version one and version two on either the end-of-semester test or on the retest.

Discussion

The three categories of information which were lost, using the quadratic formula, completing the square, and combining exponents of common terms, all require the subject to remember a specific rule. On the other hand, categories which did not exhibit forgetting, such as simplifying expressions and order of operations, are sub-procedures involved in manipulating equations. These procedures are likely to be used together, and they may be more well-integrated than those procedures involving the use of isolated rules. The differences in integrability of procedures may account for the different retention patterns between these two types of information. However, it may also be that the procedures used in manipulating equations are practiced more than those involving the use of specific rules. At the present time, these two hypotheses cannot be discriminated with respect to accounting for the retention differences.

The best predictor of percent correct on the retention test was the final grade which the student received for the algebra course. Bahrick (1984) found that course grades as well as level of training in Spanish significantly predicted the scores on the retention test. In the current study, the level of training of all subjects was fairly equivalent, since the retest was given near the beginning of the semester following the completion of their algebra

course. Differences in level of training might be indicated by whether or not the student was taking a mathematics course at the time of the retest. However, this factor did not significantly contribute to predicting retest score. But again, the retest was given during the beginning of the semester following the Spring 1987 semester, so even students who were currently taking a mathematics course may not have learned much beyond what they had learned in the previous algebra course. Nevertheless, whether or not the student was taking a mathematics course at the time of the retest did not significantly predict retest score. Perhaps this is also due to having a short retention interval (relative to those examined by Bahrick).

In the next study, students were given a pretest before beginning the algebra course in order to determine which categories of items were well-learned and which were not before the course began. It was intended to determine the type of items that students were weak at before they took the course, and to determine if categories of items which were not well-learned at the start of the course were more susceptible to forgetting than those categories of information that students knew before the course began.

Study 2

Method

Subjects

The subjects consisted of students who were enrolled in the same algebra course and with the same instructor as the previous group, but who were taking the course in the Fall semester of 1987. Ninety-three of these students were given a pretest at the beginning of the semester. The score for this test was not counted toward the student's grade. The instructor explained to the students that this initial test would be used to give him an idea of the students' prior knowledge of algebra.

Seventy-five of these subjects were also tested at the end of the Fall 1987 semester. A total of 111 subjects were tested at the end of the Fall 1987 semester. The score for the test at the end of the semester was counted toward the student's grade.

Materials

The same two versions of the algebra test were used as in the previous study.

Procedure

The pretest consisted of version one of the algebra test. The students were given 50 minutes to complete this test. The test was given during one of the students' regular class periods. The test given at the end of the semester was version two of the algebra test. Before starting the test, the students were requested to fill out a form to indicate their current and permanent addresses and phone numbers. They were informed that they would be contacted at a later date to return for testing. They were also told that if they chose to be retested, they would receive payment for their time. Students were given 50 minutes to complete the test. The test was given during a regular class period.

Results

Students who took both the pretest and the end-of-semester test (N=75) showed a significant increase in percent correct from the pretest (Mean=72.4) to the end-of-semester test (Mean=84.1), ($t=11.0$, $p<.01$). There was no significant difference in end-of-semester test scores between the students who had taken the pretest (Mean=84.0, N=75), and the students who had not taken the pretest (Mean=84.0, N=36).

There were several categories of items on which students did poorly on the pretest. In Table 2, the categories of items on which less than 50% of the

students provided correct answers are indicated by asterisks (also see Table 3 for percent correct by category on the pretest). The number of asterisks indicates how many of the items were below 50% correct. Students had difficulty with three categories of items on the pretest: using the quadratic formula, finding a common denominator, and simplifying expressions.

Discussion

Before the algebra course began, subjects were weak at solving problems in a number of categories: using the quadratic formula, finding common denominators, and in most of the procedures used in manipulating equations. Although different groups of students participated in Study 1 and in Study 2, a comparison of those categories of items which students had difficulty with on the pretest (Study 2) and those categories of items which showed a retention loss on the retest (Study 1) may be suggestive. Upon making this comparison, there is no clear evidence that categories of items which students performed poorly on in the pretest were those categories which were likely to be forgotten (See Table 2). For example, pretest students did poorly on the quadratic formula problems, and retested students showed forgetting on these items. On the other hand, pretested students also performed poorly on items in which they had to simplify expressions, but retested students did not show forgetting of this category of items. These comparisons are only meant to be suggestive; the students from Study 2 are currently being retested, and the results from that test, rather than a comparison across different groups, will provide evidence for or against the hypothesis that information not known before the algebra course was taken is more susceptible to forgetting than information which was known before the course began. It is also hoped that a larger sample of students will be retested than in Study 1.

Further Study

As stated earlier, the students from Study 2 are currently being retested. We also hope to re-design the algebra test in order to obtain a better understanding of the categories of algebra knowledge and skills which are forgotton or retained after some period of disuse.

Acknowledgements

The authors thank Clayton Lewis for providing sources and references to material about students' understanding of algebra.

References

Bahrick, H.P., Bahrick, P.O., & Wittlinger, R.P. (1975). Fifty years of memories for names and faces: A cross-sectional approach. Journal of Experimental Psychology: General, 104, 54-75.

Bahrick, H.P. (1983). The cognitive map of a city- 50 years of learning and memory. In G.Bower (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, 17, 125-163. New York: Academic Press.

Bahrick, H.P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. Journal of Experimental Psychology: General, 113, 1-29.

Bundy, A. (1975). Analysing Mathematical Proofs (or reading between the lines). Unversity of Edinburgh, Department of Artificial Intelligence Research Report 2.

Carry, L.R., Lewis, C., & Bernard, J.E. (no date). Psychology of equation solving: An information processing study. Technical Report, Department of Curriculum & Instruction, Unversity of Texas at Austin.

Matz, M. (1982). Towards a generative theory of high school algebra errors. In D. Sleeman & J.S. Brown (Eds.), Intelligent Tutoring Systems. London/New York: Academic Press.

Sleeman, D. (1982). Assessing aspects of competence in basic algebra. In D. Sleeman & J.S. Brown (Eds.), Intelligent Tutoring Systems. London/New York: Academic Press.

Sleeman, D. (1984). An attempt to understand students' understanding of basic algebra. Cognitive Science, 8, 387-412.

Svokowski, E.W. (1986). Fundamentals of Algebra and Trigonometry, Sixth Edition. Boston: Prindle, Weber, & Schmidt.

APPENDIX F

A COMPONENTIAL ANALYSIS OF THE KEYWORD METHOD

Robert J. Crutcher and K. Anders Ericsson
University of Colorado at Boulder

Introduction

The keyword method is a two-step mnemonic to learn new vocabulary. First, a new vocabulary word is related to a keyword, a concrete, imageable English word that is acoustically similiar to the new vocabulary word or to a salient part of the new word. Then, an interactive image between the keyword and the English equivalent for the new vocabulary word is created. For example, to learn the Spanish word 'doronico,' which means 'leopard,' one might link the word 'door' (keyword) with 'doronico' and then create an interactive image using'door' and 'leopard' (e.g. an image of a leopard springing through a door). There has been considerable research on the keyword method, much of it attempting to demonstrate the superiority of the method to other vocabulary learning techniques (e.g. Atkinson & Raugh, 1975; Raugh & Atkinson, 1975; McDaniel & Pressley, 1984). Though some studies have tried to identify individual differences (e.g. age and ability) that might influence the method's effectiveness (e.g. Delaney, 1978), few studies have focused on differences in processing steps or components.

The goals of the study reported here were: (1.) To develop a methodology that would provide a detailed description of the encoding and retrieval processes involved in the keyword method; (2.) To use this methodology to decompose the retrieval processes involved in the keyword method; and (3.) To identify any differences in the retention characteristics of the various components of the keyword method and the implications of such differences for teaching the keyword method. In keeping with these goals, we collected three types of data: cued recall, retrieval times, and verbal reports. For the cued recall tasks, we tested each of the component tasks (i.e. the Spanish to keyword component and the keyword to English word component) as well as the overall task (i.e. the Spanish to English word task). Retrieval times for the overall task as well as the component tasks were collected to provide an alternative and possibly more sensitive measure of retention. Finally, verbal reports were collected to provide a detailed record of the encoding and retrieval steps. .The verbal reports are not fully analyzed as of yet and are therefore not discussed here.

Method

Subjects. The subjects were 24 undergraduate students enrolled in an introductory psychology course at the University of Colorado, Boulder.

Design and Procedure. Subjects were intially assigned to either a 1-week or 1-month retention group. After instructions in how to give verbal reports (Ericsson & Simon, 1984) and in how to use the keyword method, all subjects learned a list of 42 Spanish vocabulary items using the keyword method. The items were presented on an IBM-PC with the Spanish word on the far left, the keyword in the middle, and the English word on the far right. Presentation was self-paced with a maximum study time of 20 seconds per item. For half of the items, subjects were asked to think-aloud while studying. Think-aloud and silent items were counterbalanced across subjects.

To ensure that all items were learned, following the study phase was a dropout phase, in which subjects were tested on all vocabulary items to one correct retrieval. Subjects saw the Spanish word and were required to say the English word into a microphone, which was connected to a voice relay that recorded the response latency. Feedback, consisting of the original three words (i.e. Spanish word, Keyword, and English word), was provided after each incorrect response.

Following the dropout phase, all subjects were tested on three retrieval tasks: (1) Full Retrieval Task: Given the Spanish word, recall the keyword; (2) Keyword Retrieval Task: Given the Spanish word, recall the keyword; and (3) Image Retrieval Task: Given the keyword, recall the English word. Retrospective reports were taken after those items that subjects had studied thinking aloud. For each test, which consisted of 42 trials, a third of the trials were of each retrieval task type, so that three tests of 42 trials each were required to test all 42 items on all 3 retrieval tasks. Items and retrieval tasks were counterbalanced as well as the 6 possible retrieval task orders. Subjects completed two blocks of 3 tests each. Half the subjects were retested after 1 week and the other half after 1 month.

Results and Discussion

All of the results reported are for the second test block at immediate and the first test block at delay.

Cued Recall. For the full retrieval task (Spanish to English), mean recall dropped from 97% to 86% for the 1-week group, a small but reliable loss, and from 94% to 37% for the 1-month group, a large and significant loss (Figure 1.).

Recall results for the component retrieval tasks are compared to those for the full retreival task in Figures 2 and 3. For the keyword retrieval task, mean recall was 98% at immediate and 95% at delay for the 1-week group, not a significant difference, and 95% at immediate and 88% at delay for the 1-month group, a small but reliable difference. However, for the image retrieval task, mean recall declined from 97% to 88% for the 1-week group, a small but reliable loss, and from 96% to 42% for the 1-month group, a large and significant loss.

This pattern of results, the keyword showing essentially no retention loss after either delay interval, but the image and full retrieval tasks showing parallel retention losses (9% and 11% after a week and 54% and 57% after a month), suggests that the decline in overall retention was due to a loss associated with the image component and not the keyword component. In other words, subjects almost invariably recalled the keyword given the Spanish word, but could not always recall the English word given the keyword.

Recall Latencies. There was a significant increase in retrieval times for the full retrieval task after a week: from 1815 msecs to 2368 msecs, an increase of 553 msecs (Figure 4.). After a month, the increase was even greater: from 1855 msecs to 3216 msecs--a difference of 1361 msecs.

In Figures 5 and 6 latencies for the component tasks are compared to those for the full retrieval task. For the keyword task, there was a small but reliable increase in retrieval time after a week: from 1427 msecs to 1709 msecs, an increase of 282 msecs; and a much larger,

highly significant increase after a month: from 1481 msecs to 2206 msecs, an increase of 725 msecs. For the image retrieval task, there was a significant increase in retrieval time after 1 week: from 1471 msecs to 2000 msecs, an increase of 529 msecs; and a much larger and significant increase after 1 month: from 1556 msecs to 2768 msecs, an increase of 1212 msecs

The latency results agree quite well with the recall results, though clearly they provide a much more sensitive index of retention loss. Again, the pattern of results suggests that the image component decays more readily than the keyword component: retrieval-time increases for the image component are approximately twice those for the keyword component at both delay intervals. Moreover, the increases in latencies for the image component parallel the increases in latencies for the full retrieval task, again suggesting (as did the recall data) that the decay in the overall retrieval task is due to the image component.decaying. However, it should be noted that whereas the recall data showed no significant retention loss for the keyword component after a week and only a small loss after a month, the latency data show a significant loss retention loss for the keyword after even a single week.

It is interesting to compare the performance results for the image component after 1 week with those of the keyword results after 1 month. The increase in latency for the image component after 1 week was 529 msecs; for the keyword component after 1 month, 725 msecs. The decrease in recall for the image component after 1 week was 9%; for the keyword component after 1 month, 7%. There is the possibility that retrieval times might be a reliable predictor of recall performance.

Summary

The results discussed in the present paper suggest that the two components of the keyword mnemonic decay differentially, with the image component decaying more rapidly. The practical implication of these results is that it may pay to practice the image component more or to find more distinctive encodings to relate the keyword and English word. However, as with most keyword studies, the keywords selected in this study were quite similar to the Spanish words. Using less-similar keywords might alter the results found here.

Finally, the present results argue for the importance of multiple measures as well as the usefulness of a componential analysis. Retrieval latency may be a more sensitive predictor of retention loss than recall; in fact, it may prove a useful predictor of subsequent recall performance. In additional analyses, not reported here, we have found that recall latency for the full retrieval task at immediate test reliably predicts recall performance after delay.